ADVANCES IN GENOME-SCALE MODELING APPLIED TO EXPRESSION-BASED FLUX ESTIMATION AND EPISTASIS PREDICTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Brandon Elam Barker

August 2014

ADVANCES IN GENOME-SCALE MODELING APPLIED TO EXPRESSION-BASED FLUX ESTIMATION AND EPISTASIS PREDICTION

Brandon Elam Barker, Ph.D.

Cornell University 2014

Quantitative models are increasingly being used to interrogate the metabolic pathways that are contained within complex biological processes, and at a higher level, these models are used to explore questions in evolution with complex physiological processes absent in typical, idealized population genetic models. In this work, we focus both on the application of quantitative models in evolution and the development of new quantitative methods for metabolism. An overview of constraint-based modeling and its purview in the field of metabolic modeling is given in Chapter 1. By using a simple version of constraint-based modeling known as flux balance analysis (FBA), we elucidate patterns that occur in gene-gene interactions of deleterious mutations (Chapters 2 and 3). Because many biological problems relate to systems that are not well-suited to FBA, especially when establishing a physiologically accurate flux is desirable, we address the problem of estimating metabolic fluxes using constraint-based models and readily available gene expression data by developing a new methodology and software (called FALCON; Chapter 4). We then take advantage of the FALCON method by using it in the development of approaches that enable the simulation of beneficial mutations and reveal some of the influences that metabolic networks bring to bear on the study of adaptation (Chapter 5).

BIOGRAPHICAL SKETCH

I'd always been interested in nature since I grew up in a rural area, and this later transitioned into interest in biology. Along the way, I also became interested in computing due to my exposure to them through my parents' business, which encouraged me to focus on mathematics and programming.

I picked up biology again more seriously near the end of my undergraduate studies and began working as a bioinformatics cluster administrator, which soon transitioned into a full time position that included maintaining a database, training its users, and extending its features. I developed a novel, parallel algorithm for biological sequence alignment, conducted comparative genomics research that relied heavily on the UNIX style of data processing, and I initiated collaboration with a statistician on tree of life problems.

My graduate studies have focused on optimization and mathematical modeling, in particular, their applications to the modeling of metabolism and simulation and analysis of epistasis.

To my parents—James and Carol Barker,

For their love, time, and beneficence,

Without which this work would not have been possible.

To my wife, Lin Xue,

For giving me the momentum to pursue graduate school,

For her patience while I studied,

For her continuing love,

And most importantly, for being an amazing mother.

To my son, Lindon Barker,

For your extra motivation towards finishing in a timely manner,

And for your terrific smiles, grins, and laughs during the last two years.

ACKNOWLEDGEMENTS

Firstly I would like to thank my advisor, Zhenglong Gu, under whose supervision this work has unfolded, and who has provided me with substantial support, encouragement, and irreplaceable mentoring. I owe much to my colleague and friend, Lin Xu, for not only showing me the ropes and giving me hope, but also getting me interested in so many topics in evolution; much of the present work, particularly that involving epistasis, would not have happened without him. Without Jason Locasale's intellectual incitation and advice, I never would have wandered down the path to what has, to me, become the most interesting part of this work, nor the path to the Big Red Barn quite so much as I did. My thesis committee members—David Christini, Chris Myers, and Michael Stillman—have provided much advice over the years, scientific and otherwise. Without Michael's advice to investigate FBA when I started graduate school, it is very likely my life, and this work, could have been very different.

My other collaborators in the present work include Tim Connallon, whose knowledge on epistasis surpasses anyone I know, and similarly for Alex Shestov with regard to metabolic flux analysis; Kieran Smallbone, who has given helpful advice over the years regarding constraint-based models, as well as having helped develop several algorithms and models that I have employed in this work, including one in particular that served as the impetus for some of the research in the latter chapters; Yiping Wang, a phenomenal young researcher who helped with some analysis in the last chapter during his first month on campus, and Hongwei Xi, for providing substantial support for his programming language ATS, which was employed in the latter chapters.

In the lab, Xiaoxian Guo, Huifeng Jiang, and Zhe Wang provided much experimental advice, training, and data.

I would also like to thank many other colleagues for their help, friendship, scientific discussion, or some combination of the three. These include Diana Chang, Elijah Bog-

art, Hong Chian, Martha Field, Lei Huang, Haley Hunter-Zinck, Ben Heavner, Xiaojing Liu, Henry Lu, Anuttama Mohan, Lonnie Princehouse, Narayanan Sadagopan, Marc Warmoes, and Kaixiong Ye.

Without the stimulus, confidence, and aid from my early mentors at the University of Kentucky, this document would not exist; in particular, I thank Ruriko Yoshida, for renewing my passion in science, Jim Lund, for being my first biology mentor (and providing the opportunity to become acquainted with my wife), Jerzy Jaromczyk for his friendly help and advice throughout and beyond my undergraduate education, and to Henry Dietz, for being a great educator and advising my senior design project. My passionate and reassuring high school biology teacher, David Christiansen, warrants special thanks for being continually amazed at the wonders biology has to offer. Lastly, I thank my friend Benjamin Runyon, for always lending an ear.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

There has been a surge of interest in understanding the regulation of metabolic networks involved in disease in recent years. Quantitative models are increasingly being used to interrogate the metabolic pathways that are contained within this complex disease biology. At the core of this effort is the mathematical modeling of central carbon metabolism involving glycolysis and the citric acid cycle (referred to as energy metabolism). Here we discuss several approaches used to quantitatively model metabolic pathways relating to energy metabolism and discuss their formalisms, successes, and limitations [1].

The accumulated amount of biochemical work carried out over the years has elaborated complex metabolic systems and networks. This information includes the network architecture encoded in chemical reactions that are carried out by metabolic enzymes and the kinetic parameters that determine reaction mechanisms involved in each of these chemical reactions. Application of this knowledge has led to tremendous predictive capability in characterizing metabolic regulation in normal physiology including the growth of unicellular organisms and the successful simulation of energy metabolism in healthy red blood cells. However, there are far fewer instances in which these models have been applied to the characterization of pathophysiology. Applying our knowledge of metabolic regulation to the investigation of disease states such as cancer or neurodegeneration is currently a scientific frontier. In this review, we will revisit several classic techniques for the mathematical modeling of metabolic pathways and discuss instances

[1] This chapter is taken from material in Shestov et al. 2. Brandon Barker is the primary author of all material found herein.

where their application to biomedical science is beginning to yield fruitful dividends.

1.1 Linear Systems, Flux Balance Analysis

Linear models are mathematical models that contain a set of algebraic equations based on the stoichiometric relationships that define conservation relationships within a metabolic network. Linear models, to our knowledge, were first applied to biochemical systems in 1961 by Howard Shapiro [10]. Shapiro discussed the possibility of using optimization in biochemical linear models in a 1969 publication [11]. In 1984, a model incorporating glycolysis and the TCA cycle was employed running a variant of Dantzigs algorithm with the assumed biological objective of minimized free energy dissipation [12, 13]. An enduring research program was initiated by Bernhard Palsson half a decade later [14, 15]. One of Palsson's early works showed that growth maximization in an *E. coli* model could correctly match 86% of 79 gene essentialities examined [16]. Subsequent modeling in *S. cerevisiae* was able to closely predict growth rates and exometabolic fluxes in various media, and nearly capture the in vivo phosphate/oxygen (P/O) ratio of 0.95 with a simulated P/O value of 1.04, showing that models of eukaryotes were also feasible [17]. If one chooses the biological objective function to reflect the appropriate physiological demands then it is possible to predict features of adaptation; this was shown to be the case for growth optimization in several E. coli mutants [18]. By this time it had become apparent that linear models held much promise, particularly when coupled with optimization.

1.2 Genome Scale Modeling

Today, when we refer to linear models, we most often mean Constraint Based Models (CBMs). We refer to a CBM as any model making use of the stoichiometric matrix, **S**, as a linear matrix constraint, e.g. $\mathbf{SF} = \mathbf{0}$, where **F** is a flux vector[2]. In fact, this is a nearly universal constraint, as it guarantees conservation of mass during steady state processes such as exponential growth or tissue maintenance [19]. Other constraints commonly used include reversibility constraints when the direction of a reaction is known for physiological conditions of interest, bounds on the uptake of nutrients or efflux rates due to regulation or physiology, or bounds on enzyme reactions when the maximum enzyme velocity V_{max} is known.

Because these constraints give rise to an underdetermined system, it will not be possible to identify a unique solution for the flux vector. A unique solution is often desirable as it allows investigators to analyze a putative metabolic phenotype. Indeed, this is one of the more convenient features of linear optimization: the ability to get meaningful solutions without explicitly taking into account any, or at least very few, free parameters. Flux Balance Analysis, or FBA, assumes a linear combination of fluxes to be maximized or minimized (Figure 1.1). In microbes, perhaps the most popular FBA objective has been growth maximization, which consists of the biomass precursors and products formulated as a single pseudo-reaction. Additionally, an ATP maintenance constraint should be formulated as a sink reaction with the molar ATP required to keep one gram of dry weight biomass living for one hour [20]. This empirically determined constraint, although assumed, is less discussed, perhaps due to its dependence on individual strains and environments. We note that for many expression-based methods in the CBM framework, the ATP maintenance constraint is not required (see Table 1.1 and

[2] F is usually used to denote a steady state flux, whereas v is used for fluxes more generally. In much of the genome-scale constraint-based modeling literature, v is used in all cases

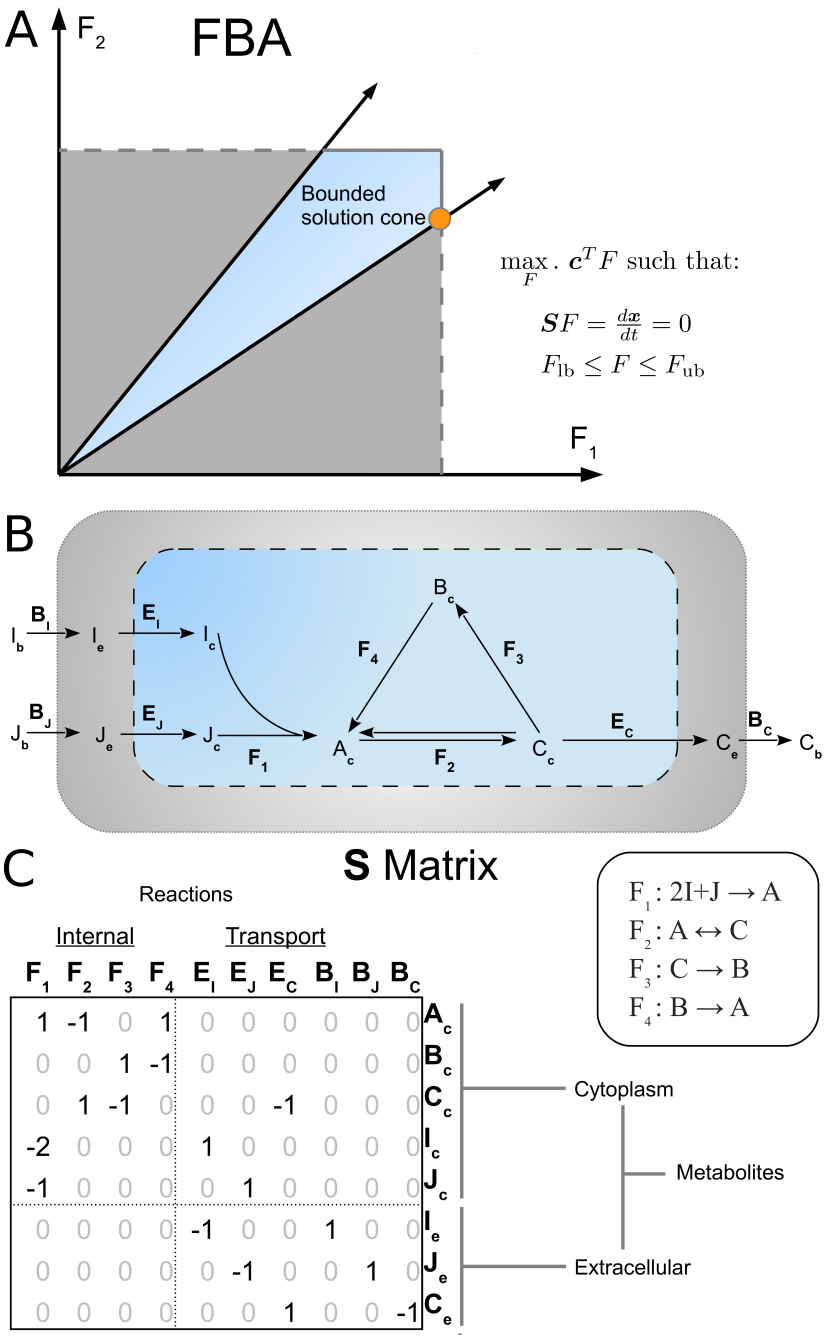


Figure 1.1: A simple geometric illustration of an FBA problem **(a)**. Constant constraints on the F_i limit the feasible solution to an n-dimensional cube (shown in grey). Further linear constraints from the **S** matrix create a cone of feasible solutions (blue). Linear programming algorithms find an optimal solution on a vertex (illustrated with orange circle). Depiction of a simple metabolic network with compartmentalization and its associated stoichiometric matrix **(b, c)**. The three compartments denoted with subscripts *b*, *e* and *c* represent the boundary, extracellular environment, and cytosol, respectively. The boundary is what separates the model from its environment, and mass balance is not assumed at the boundary; this allows for the implementation of source and sink reactions.

Figure 1.2 for examples). Fixed biomass objectives by themselves also have some undesirable qualities; biomass composition likely has some measure of variability based on genetic background and environment. Robust FBA attempts to address this problem by allowing some variation in the biomass composition, as determined by variation of empirical assays of biomass [21]. Despite these caveats, FBA has recently been found to not only predict growth in microbes, but also has good agreement with gold standard 13C flux assays in vivo when the growth objective is used along with ATP synthesis maximization and minimization of absolute fluxes [7].

Minimization of absolute flux is a commonly used objective employed alongside other objectives, forming a minimax problem (i.e. finding the minimum absolute flux profile among all flux profiles that maximize biomass). This approximates the biological goal of being efficient with enzyme production costs and enzyme crowding constraints while also guaranteeing that no thermodynamically impossible loops are present, that is, ruling out some fluxes that might otherwise violate Kirchoffs loop rule [8, 25]. This constraint will work whenever a sink reaction, such as growth, is being optimized. However, maximizing an internal flux, as in Flux Variability Analysis [22], could still result in internal cycles [25]. Initial thermodynamic approaches involved nonlinear optimization [26, 27, 35, 36]. Constraints satisfying Kirchoffs loop rule were later developed that were faster and more generally applicable than prior methods [25, 37]. Still, these involve integer constraints that put this problem in a slower class of algorithms than the convex minimized absolute flux problem. When available, thermodynamic data is valuable; it can not only be used to guarantee there are no internal cycles, but can also aid in determining reaction direction and potential regulatory targets [25, 35, 38, 39]. Application of this framework to concentration data allows unmeasured metabolite concentrations to be inferred and global concentrations to be resolved at the organelle level [36]. CBMs have also found use in tracing individual atoms through pathways, which

Table 1.1: Families of methods for constraint-based models. Broad classes of methods are described, along with references to some individual implementations or studies.

Method Family	Description	Benefits	Caveats	Solver Type	Notes
FBA [22]	Flux Balance Analysis: Linear programming applied to the model.	Usually very fast and simple to use, especially when a biomass pseudo-objective is available.	Arguably has more limited use in non-microbial models. Only simple objectives or sequential (e.g. bi-level) optimization is practical.	Linear	Often constraint-based modeling (CBM) in general may be referred to as FBA, though this is not technically correct.
MoMA [23, 24]	Minimization of Metabolic Adjustment	Usually very fast and simple to use, especially when a reference or wild-type flux is available; useful for simulating mutations.	It has been argued that the closest distance to a flux doesn't represent mutation as well as simulating the least number of flux changes (ROOM).	Linear, Quadratic Convex	Related, but slightly more sophisticated methods are being used to estimate flux profiles from expression data.
DFBA [20]	Dynamic FBA: incorporates a step-wise simulation of FBA, along with update rules that relate biomass to uptake rate, solving for extracellular concentrations.	Allows for some non-steady state observations	Small timescale dynamics and intracellular dynamics may be difficult to model.	Linear (Iterative)	Other, but infrequently used (due to difficulty) methods involving regulation (rFBA) or multi-scale models of tissues build on this approach.
EBA [19, 25, 26, 27]	Energy Balance Analysis: FBA, but also incorporates thermodynamic constraints	Incorporates thermodynamic information, prevents futile cycles.	Usually much slower than LP methods like FBA.	nonlinear, MILP, or Monotropic	A highly active research area.
Tissue-specific Model Creation [28, 29, 30]	Requires expression data for tissue of interest.	Tissues have vastly different regulatory schemes; these methods take this into account by finding which metabolic genes are likely to be expressed in a given tissue.	Still requires some other method and objective to estimate flux or do pathway analysis.	MILP	A highly active research area.
Expression-Flux mapping [1, 31]	Takes ideas from MOMA and tissue-specific model creation to estimate fluxes.	Unlike tissue-specific models, will actually estimate the flux since a MOMA-like objective is employed.	Requires high-quality (e.g. RNA-Seq) expression data, or for PROM, abundant microarray data from different conditions.	Linear optimization, but moderate number of simulations or preprocessing required.	Highly accurate predictions can be obtained.
Interaction Search [9, 32, 33, 34]	Epistasis, or genetic interactions, come up in many contexts, but are also important in energy metabolism, since energy is often related to very important phenotypes including growth, proliferation, and survival.	For such analyses, convex optimization may offer the only tractable method.	Simulating pairwise epistasis in the general case requires pairwise simulation of all double mutants of interest, which can be very time-consuming at the genome scale when different mutations in each gene, or different environments, are considered	Linear optimization, but often many simulations required. Min Cuts (exponential).	The sign of weak epistasis is difficult to predict, due to error propagation in growth rates.

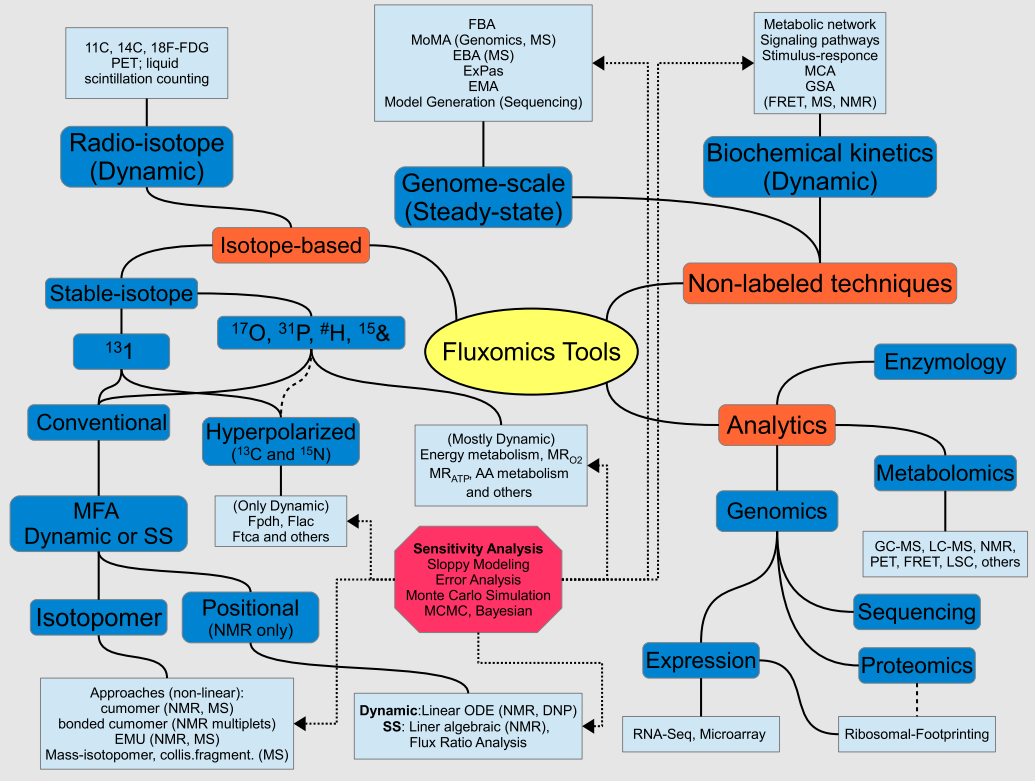


Figure 1.2: Schematic representation of fluxomics tools. Important to fluxomics are both the mathematical and computational tools for nonlabeled and labeled techniques, as well as the analytical methods used to obtain data and parameters. In the current work we focus only on nonlabeled genome-scale steady state and related analytical methods, but a full description can be found [2]. Sequence data is employed in the construction of organism models, whereas proteomics and expression data find use in the creation of tissue or cell-type-specific models. High-quality expression data such as RNA-seq and ribosomal footprinting are beginning to find uses in flux prediction. Several prominent genome-scale techniques include flux balance analysis (FBA), minimization of metabolic adjustment (MoMA), energy balance analysis (EBA), ExPas (extreme pathways), and elementary mode analysis (EMA). Nonlabeled techniques along with genome-scale analysis include biochemical kinetics modeling tools to study metabolic and signaling networks and their regulation architecture with established tools like metabolic control analysis (MCA) and global sensitivity analysis (GSA). Additional sensitivity analysis should be conducted, e.g., with Monte-Carlo techniques like Markov chain Monte-Carlo (MCMC, Bayesian) analysis to check the reliability of extracted metabolic parameters, including fluxes.

provides a more appealing framework for performing Metabolic Flux Analysis (MFA; discussed below) on stable isotope data due to the lack of bias compared to typical MFA models, which are often an order of magnitude smaller than genome-scale reconstructions [40]. Recent insightful work has made it possible to simplify the computational complexity of loopless FBA to be nearly the same as conventional FBA, but some mathematical difficulties must still be overcome before bounds on exchange fluxes can be suitably incorporated for genome-scale modeling [19, 41].

The metabolism of different tissues within the same organism is diverse; whereas the metabolism in liver is anabolic, neurons or red blood cells have a much more limited catabolic regime [28, 42, 43].The creation of tissue specific models for multicellular organism has become an important problem, and several automated algorithms taking as inputs tissue expression data and a generic model for the organism have been developed [28, 29, 30]. Coupling multiple cellular models together will enable multi-scale modeling of tissues in multicellular models or entire ecosystems for microbes [42, 44, 45, 46].

Automated generation of metabolic networks from genome sequence and pathway databases, especially in prokaryotes, has been developed [47, 48, 49, 50]. This will offer many advantages to modelers: a starting point for curated models (a draft reconstruction is estimated to often take several months even in prokaryotes), a means for doing population or ecological simulation [46], and personalized genomic modeling for patients with metabolic syndromes such as cancer where both the patient and possibly the disease have diverse genotypes [51, 52]. Eukaryotic models are somewhat more difficult to generate due to the necessity of protein localization and metabolite transporter information [47]. Automatic reconstruction going beyond enzymatic gene information, such as rFBA models, should also be possible [53, 54]; the automated generation of Boolean and higher-order discrete regulatory models using time-series expression data

has been explored as well, though to date these regulatory models have not been coupled to metabolic reconstructions [55, 56, 57, 58]. These approaches and other families of genome-scale methods are discussed in Table 1.1.

Several approaches have been used in applying CBMs to cancer and the Warburg effect, the preference for glycolytic ATP production over glucose-derived mitochondrial ATP production in cancer cells [59, 60, 61]. An important study working with a simplified, small model of central-carbon metabolism showed that, while the TCA cycle predicts better ATP yield than glycolysis when only available glucose is considered as a constraint, the addition of enzyme solvent-capacity constraints creates a preference for ATP synthesis through glycolysis [61]. More recently, the work of Vazquez et al. was extended to include a genome-scale model along with enzyme solvent-capacity constraints, which was able to show significant correlations between fluxes and expression in the NCI-60 cell line panel, as well as predicting an intermediate state in cancer metabolism transition exhibiting a temporary increase in OxPhos that was supported by two prior experimental observations [60]. All of these approaches correctly predicted lactate production. Concurrent research on predicting cancer targets by screening for simulated negative epistasis in cancer tissue-specific models that have at least one known-drug target and no known effect on normal tissue revealed many epistatic interactions [52]. A related study confirmed one of these synthetic lethalities between hemeoxygenase and fumarate hydratase, a mutation found in certain kidney cancers [62]. The recent publication of Human Recon 2 promises to aid in the understanding of many human diseases; already 65 cell-type specific models based on it are available, and the model reports 77% accuracy in identifying metabolic markers across 49 inborn errors of metabolism [63]. Although this model is a great step forward in consolidating much of the knowledge about human metabolism, it is only one of many steps to come. For instance, this model is still primarily only amenable to steady-state approaches,

lacks corresponding enzyme-regulatory and signaling architecture, and has introduced more dead-end metabolites than it removed (1,176 versus 339).

1.3 Conclusions for the State of Linear and Genome-Scale Models

Kinetic models for smaller pathways are possible when the data are present, but many energetic questions concern the entire cell, leaving only incorporation of CBMs as a viable option. The original efficiency and ease of use of FBA have helped propagate a field of more diverse algorithms that are often tractable on todays computers using the same modeling and software frameworks [64, 65]. Numerous methods and successful applications in energy metabolism exist, including prevalent diseases such as heart disease, cancer, and Alzheimers [66].

Multiscale models, as were used in the Alzheimers models, will undoubtedly become more common. At the intracellular scale, CBMs are also beginning to incorporate information other than metabolic stoichiometry [67, 68, 69]. A whole cell model for *Mycoplasma genitalium* incorporating information about all classes of macromolecular synthesis and degradation, in addition to stoichiometric and regulatory information, found a non-stochastic coupling between metabolism and the cell-cycle where DNA replication rates depended on the concentration of dNTP [68]. Models like these are not easy to build, but substantial endeavors are underway to assist in their draft construction and refinement, and together with an increase in use of jamboree meetings of organism and model experts and online collaborative tools, will likely aid in creating public models of higher quality and the understanding of many biological processes outside the traditional scope of metabolism [48, 63, 70, 71, 72, 73, 74].

CHAPTER 2

DYNAMIC EPISTASIS FOR DIFFERENT ALLELES OF THE SAME GENE

Epistasis refers to the phenomenon in which phenotypic consequences caused by mutation of one gene depend on one or more mutations at another gene. Epistasis is critical for understanding many genetic and evolutionary processes, including pathway organization, evolution of sexual reproduction, mutational load, ploidy, genomic complexity, speciation, and the origin of life. Nevertheless, current understandings for the genome-wide distribution of epistasis are mostly inferred from interactions among one mutant type per gene, whereas how epistatic interaction partners change dynamically for different mutant alleles of the same gene is largely unknown. Here we address this issue by combining predictions from flux balance analysis and data from a recently published high-throughput experiment. Our results show that different alleles can epistatically interact with very different gene sets. Furthermore, between two random mutant alleles of the same gene, the chance for the allele with more severe mutational consequence to develop a higher percentage of negative epistasis than the other allele is 50-70% in eukaryotic organisms, but only 20-30% in bacteria and archaea. We developed a population genetics model that predicts that the observed distribution for the sign of epistasis can speed up the process of purging deleterious mutations in eukaryotic organisms. Our results indicate that epistasis among genes can be dynamically rewired at the genome level, and call on future efforts to revisit theories that can integrate epistatic dynamics among genes in biological systems[1].

[1]This chapter is published as Xu et al. [9]. Brandon Barker and Lin Xu contributed equally to this work. It is additionally available in Xu [75, chapter 4].

2.1 Introduction

Epistasis between two deleterious mutations is positive when a double mutant causes a weaker mutational defect than predicted from individual deleterious mutations, and is negative when the double mutant causes a larger defect [76, 77]. In a population with sexual reproduction, positive epistasis alleviates the total harm when multiple deleterious mutations combine together and thus reduces the effectiveness of natural selection in removing these deleterious mutations, whereas negative epistasis can lower average mutational load by efficiently purging deleterious mutants [78]. As a consequence, selective elimination of deleterious mutations would be especially effective if negative epistasis is prevalent. It is important to understand the distribution of epistasis among mutations, which plays a central role in genetics and theoretical descriptions for many evolutionary processes [76, 77].

Tremendous efforts have been put into genome-wide measurements for the sign and magnitude of epistasis among different genes in various species [79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90]. A series of high-throughput experimental platforms have been developed, such as synthetic genetic array (SGA; Costanzo et al. 79, Tong et al. 80), diploid-based synthetic lethality analysis with microarrays [81, 82], synthetic dosage-suppression and lethality screen [83, 84, 85], and epistatic miniarray profiles [86, 87, 88]. The epistatic relations in these experiments were mostly measured based on one mutant type (deletion mutant) per gene. Few studies constructed multiple mutant alleles for single genes to examine the dynamics of epistatic relations among genes under different genetic perturbations. As a consequence, the global landscape of epistasis for different alleles of the same gene remains largely uninvestigated.

We address this issue by exploring epistatic differences among alleles in the same

gene for a large part of the genome by combining experimental data with mathematical modeling using flux balance analysis (FBA). FBA involves the optimization of cellular objective functions and allows prediction of in silico flux values and/or growth [8, 91, 92]. FBA has been used to investigate the fitness consequence of single-deletion mutants [93, 94] and epistatic relations between metabolic reactions, genes, and functional modules [34, 95, 96, 97]. The FBA predictions show good agreement with genome-wide experimental studies [23, 24, 98, 99, 100, 101, 102, 103]. One essential advantage of FBA modeling is that it can simulate epistasis between genes based on different genetic mutants. Using this platform, together with data from a recently published experiment [79], we were able to show that epistasis can be rewired among genes, and that the sign of epistasis can change dramatically at the global scale, depending on the mutant alleles involved in the processes. Our study provides a genome-wide picture on the dynamic epistatic landscape of various mutant alleles for the same gene.

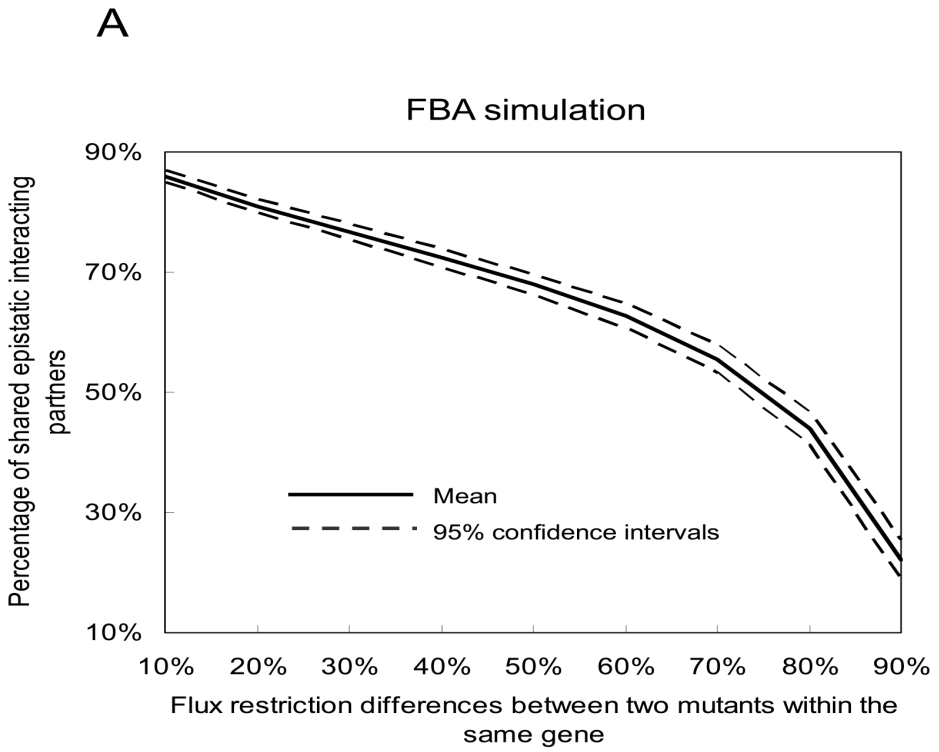
2.2 Results

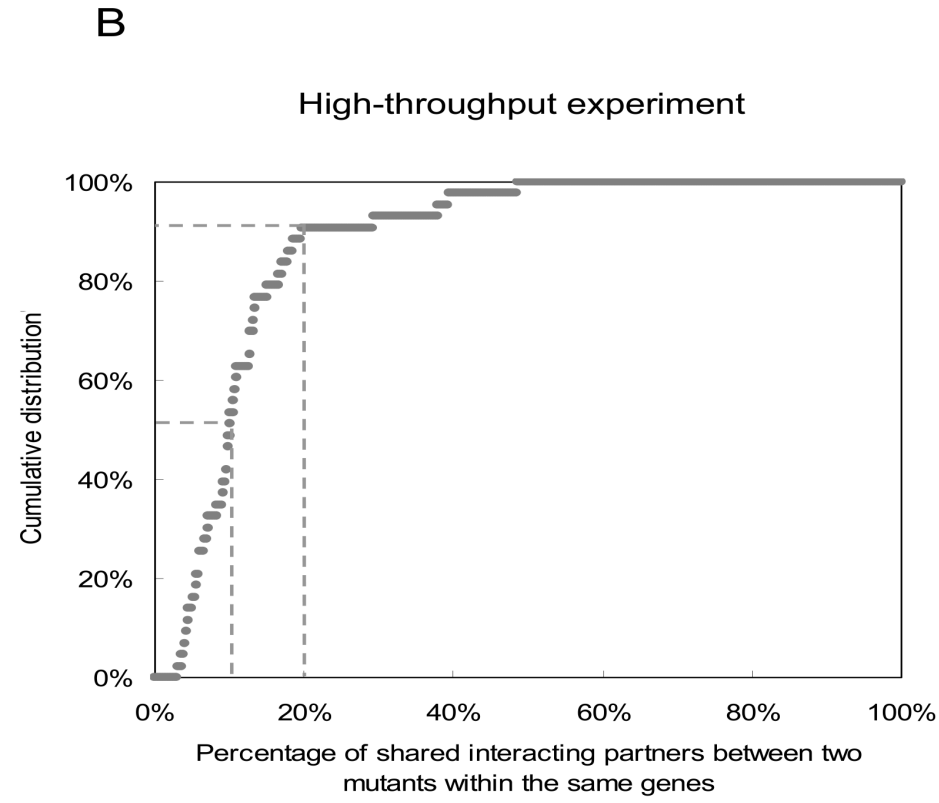
2.2.1 Epistatic Relations Between Genes Are Largely Allele-Specific.

We first used the yeast *Saccharomyces cerevisiae* metabolic reconstruction iMM904 [91] to examine the distribution of epistasis under various genetic mutant alleles. The reconstruction is a genome-scale metabolic model, having 904 metabolic genes associated with 1,412 reactions. For each gene, we simulated genetic perturbations that retain the corresponding flux from 90% to 0% in decrements of 10% of its WT (optimal) flux. As a result, 10 different single mutants per nonessential gene and nine different single

mutants per essential gene (the 0% flux mutants in these genes represent lethal deletion for which epistasis cannot be calculated) were simulated. We computed the fitness of the single mutants and double mutants with any possible pairwise allele combination of different genes. These data were used to infer the epistatic relationships among genes. In total, over 40 million simulations were conducted. To investigate the dynamics of epistasis among genes, we calculated the percentage of shared epistatic interaction partners between any two mutants within the same gene. Two mutant alleles are defined to share an epistatic interaction partner (a mutant from another gene) if they both epistatically interact with this mutant and the signs of epistasis are the same. The percentage of shared epistatic interaction partners between two mutants is calculated as the number of their shared epistatic interaction partners divided by the sum of their total epistatic interaction partners. As shown in Figure 2.2.1A, our results indicate that the percentage of shared epistatic interaction partners between two mutants of the same gene decreases as the flux difference between them increases. Two mutants of the same genes could have as low as only $\approx 20\%$ overlap between their epistatic interaction partners, indicating that the epistatic profile of a gene is largely dependent on the mutant types used. Our results also show that the average number of epistatic interaction partners per gene do not affect this conclusion (Figure A.1). Interestingly, there are cases where the sign of epistasis between two genes can even change under varying mutant types (an example is in Figure A.2, and all pairs with reversed sign of epistasis are listed in Dataset S1). However, such events are rare ($\approx 1.2\%$ of all gene pairs that show epistatic interactions). Furthermore, we repeated the above FBA analysis for another species, *Escherichia coli*, and the results confirmed the above trend (Figure A.3).

Figure 2.1: Epistatic relations between genes are allele specific. (A) FBA simulation results for the distribution of the percentage of shared epistatic interaction partners between two mutant alleles within the same gene. Solid and broken lines represent mean and 95% confidence intervals, respectively. (B) The cumulative distribution for the percentage of shared epistatic interaction partners between two mutant alleles within the same gene based on real experimental data. Two broken lines represent 10% and 20% of shared epistatic profiling, respectively.


A
FBA simulation
Percentage of shared epistatic interacting partners
90%
70%
50%
30%
10%
Mean
95% confidence intervals
10% 20% 30% 40% 50% 60% 70% 80% 90%
Flux restriction differences between two mutants within the same gene


B
High-throughput experiment
Cumulative distribution
100%
80%
60%
40%
20%
0%
0% 20% 40% 60% 80% 100%
Percentage of shared interacting partners between two mutants within the same genes

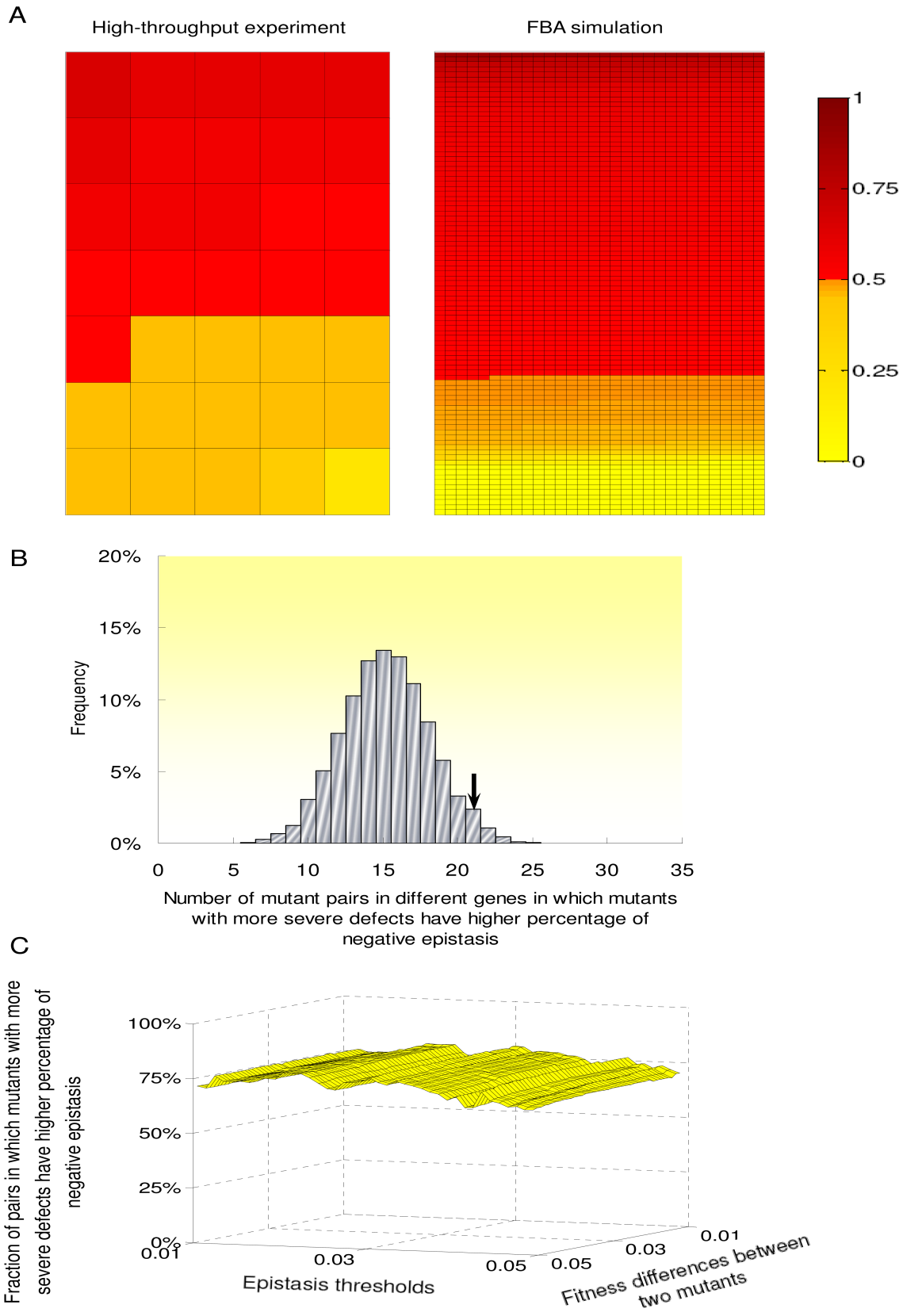
In a recently released high-throughput experiment that measured genome-wide epistatic relations among genes in *S. cerevisiae* [79], there were 43 mutant pairs having two different mutant alleles of the same gene (Dataset S2), each of which were experimentally crossed with 3,885 array gene deletion mutants to explore their epistatic relations in the genome. In total, over 200,000 double mutants were experimentally constructed. This dataset provides the most comprehensive experimental source for investigating the epistatic landscape of different mutant alleles in the same gene. Figure 2.2.1B shows the empirical cumulative distribution for the percentage of shared interaction partners between mutant pairs within the same gene. Our results indicate that more than 50% of mutant pairs within the same gene have less than 10% overlap of their epistatic interaction partners, and $\approx$ 90% mutant pairs have less than 20% overlap (Figure 2.2.1B). As shown in Dataset S2, the functions of genes used in the experiments are very diverse, and not restricted to metabolic functions as genes in the FBA model. Nevertheless, the result from experimental studies confirms our FBA modeling prediction that different mutant alleles of the same gene can have very distinct epistatic interaction partners in the genome. In addition, the conclusions are robust under various epistasis thresholds (Figure A.1).

2.2.2 Sign of Epistasis for Individual Genes Depends on Mutation Severity.

The relative prevalence of positive vs. negative epistasis is of tremendous importance for understanding many evolutionary processes [76, 77, 78]. In the following we addressed this issue for different alleles of the same gene. Based on the above high-throughput experimental dataset, we calculated the percentage of negative epistasis for each mutant,

defined as the number of negative epistatic partners for this mutant divided by the overall number of its epistatic partners. We then compared the percentage of negative epistasis between different mutant alleles of the same gene in the experiment. Among 43 mutant pairs in the study, 35 mutant pairs have significantly different fitnesses between two mutants of the same gene. As shown in Figure 2.2.2A left, 21 mutant pairs (60%) show that alleles with more severe defects have a higher chance than the other allele in the same gene to develop negative epistasis in the genome.

Figure 2.2: Mutant alleles in the same gene with more severe defects tend to have a higher percentage of negative epistasis in yeast. (A) The two matrices represent all mutant pairs identified in real experimental data (left) and FBA simulation (right) (fitness difference $|\Delta f| \geq 0.01$; epistasis threshold $|\epsilon| \geq 0.01$). Each cell represents one mutant pair within the same gene. The color bar to the right represents the normalized percentage of negative epistasis for the mutant allele with more severe defects (percentage of negative epistasis for the mutant allele with more severe defects divided by the sum of percentage of negative epistasis for two mutant alleles). Red and yellow colors represent that mutant allele with more severe defects in the same gene has higher and lower percentage of negative epistasis than the other allele, respectively. (B) Distribution for the number of mutant pairs among randomly selected 35 pairs where mutants with more severe defects have higher percentage of negative epistasis. The arrow represents the observed number for the mutant allele pairs within the same genes. (C) The percentage of mutant pairs in which the mutant allele with more severe defects in the same gene has a higher percentage of negative epistasis under various fitness difference and epistasis thresholds during FBA simulations.


A
High-throughput experiment
FBA simulation
1
0.75
0.5
0.25
0
B
20%
15%
10%
5%
0%
Frequency
0
5
10
15
20
25
30
35
Number of mutant pairs in different genes in which mutants with more severe defects have higher percentage of negative epistasis
C
Fraction of pairs in which mutants with more severe defects have higher percentage of negative epistasis
100%
75%
50%
25%
0%
0.01
0.03
0.05
Epistasis thresholds
0.05
0.03
0.01
Fitness differences between two mutants

To see if this result could be caused by a systematic trend in the high-throughput experiments, we randomly selected 35 pairs of mutants from distinct genes that have the same fitness level for single-deletion mutant and fitness difference between two mutants as the above 35 pairs of mutants within the same genes, and compared their relative prevalence of negative epistasis. The permutation was repeated 100,000 times, and the result is depicted in Figure 2.2.2B. Among all repeats of randomly selected 35 mutant pairs, only a small percentage (4.1%) have 21 or more mutant pairs where the mutant with more severe defects has a higher chance than the other mutant to develop negative epistasis in the genome, indicating that our observation for different mutant alleles of the same gene is not likely caused by the overall pattern in the high-throughput experiments.

Using results from the above FBA simulation, we also confirmed the same pattern that between mutant alleles of the same gene, the mutant allele with more severe defect is more likely than the other allele to develop negative epistasis in the genome (Figure 2.2.2A, right). Indeed, an even higher percentage of mutant allele pairs in the FBA simulation ($\approx$ 70%) than in real experiments (60%) support this conclusion. To avoid possible bias from the definition of epistasis and fitness differences between mutant alleles in the FBA simulation, we repeated the calculations based on multiple criteria and our conclusion remains the same (Figure 2.2.2C).

Our observation is surprising given that previous results based on virus models or gene network simulations proposed a totally opposite pattern at the genome level, i.e., mutations with larger mutational defects are more likely to develop positive epistasis [104, 105, 106, 107, 108]. We further used the FBA simulations to explore the dynamics of epistasis for various mutant alleles of the same gene in different species. High-quality genome-wide metabolic networks in three bacteria (*Escherichia coli* [109], *Salmonella typhimurium* [110], and *Helicobacter pylori* [111]), one archaea (*Methanosarcina bark-*

eri [112]), and another single-cell eukaryote (*Plasmodium falciparum* [113]) were used in our simulation. As shown in Figure 2.3, when two mutant alleles of the same gene are compared, in 22%, 17%, 32%, and 19% of cases for *E. coli*, *S. typhimurium*, *H. pylori*, and *M. barkeri*, respectively, mutant alleles with more severe defects display higher percentages of negative epistasis than the other allele, indicating that more deleterious mutant alleles in the same gene indeed tend to develop positive epistasis in these species. However, these numbers are significantly smaller than that of yeast and another eukaryotic organism, *P. falciparum* (52%). The conclusion is robust under various epistasis thresholds (Figure A.5).

2.2.3 Self-Purging Mechanism for Deleterious Mutations at the Population Level

Our above results indicate that between two random mutant alleles of the same gene, the chance for the allele with more severe mutational consequence to develop a higher percentage of negative epistasis than the other allele is 50-70% in eukaryotic organisms, but only 20-30% in bacteria and archaea. In other words, mutant alleles with more severe defects in the same gene might have a higher chance to develop negative epistasis in eukaryotic organisms than in bacteria and archaea. We constructed a simple population genetic model as in Figure 2.2.3A to address the evolutionary significance of this observation. The genetic system has two genes: a query gene A, which contains three different alleles (A^S: mutants with severe defects; A^D: mutants with weak defects; A^{WT}: WT), and a gene X, which has two different alleles (mutant, X^M, and WT, X^{WT}). We simulated the ratio of allele frequency between the severe and the weak mutant alleles in gene A under different probabilities of having negative epistasis between these two

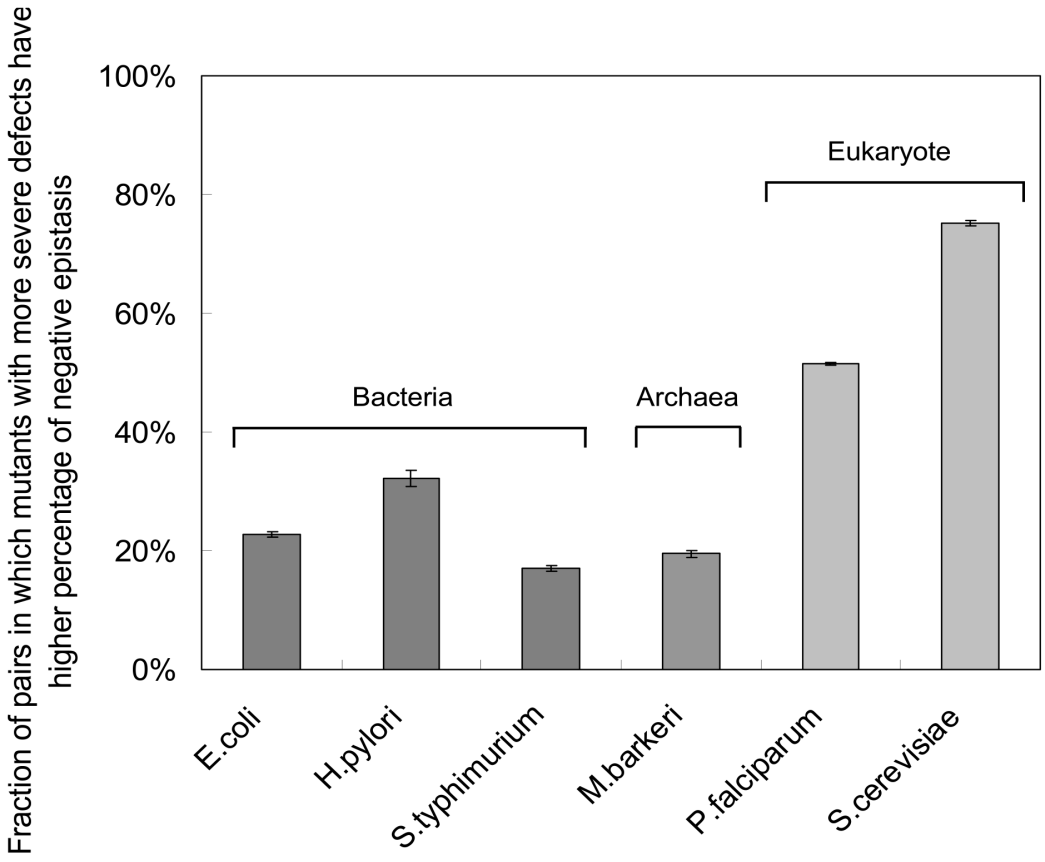


Figure 2.3: Mutant alleles with more severe defects tend to have a higher percentage of negative epistasis in eukaryotes than bacteria and archaea. The y axis shows the percentage of mutant pairs in which mutant alleles with more severe defects in the same gene have a higher percentage of negative epistasis than the other allele. FBA simulations were conducted for three bacterial species (*E. coli*, *S. typhimurium*, and *H. pylori*), one archaea species (*M. barkeri*), and two single-cell eukaryote species (*P. falciparum* and *S. cerevisiae*). The mean and SEs were based on results from 40 epistasis threshold values ranging from 0.01 to 0.05.

alleles and the mutant allele in the gene X.

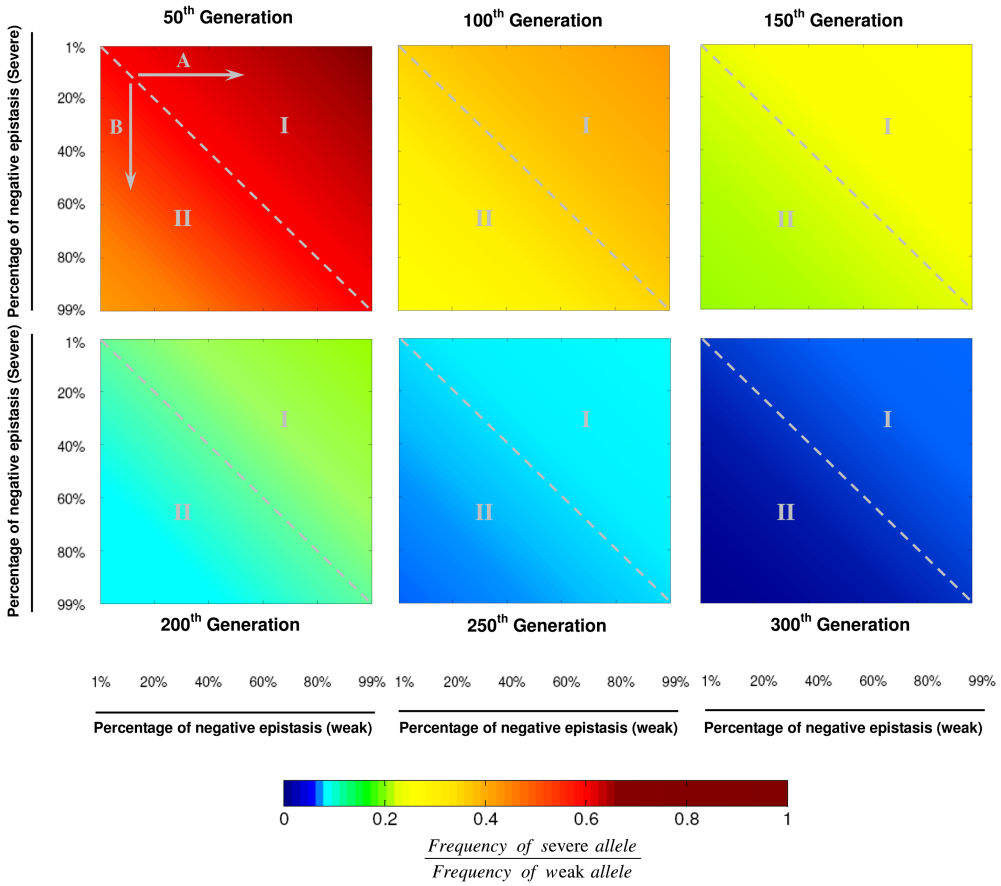
Figure 2.4: Increased efficiency of purging deleterious mutations in eukaryotic organisms. (A) The population genetics model for allele frequency changes from generation to generation. In the figure, ρ and ω represent allele frequency and fitness, respectively. A and X are genes with different alleles, and ϵ is the epistasis term between mutant types of different genes. (B) The ratio of the severe to the weak alleles of the A gene in the 50th, 100th, 150th, 200th, 250th, and 300th generations. Colors represent the ratio as indicated at the bottom. The diagonal line in each panel represents the situation where the severe and the weak mutant alleles have the same probability of having negative epistasis in the genome. It is noteworthy to point out that in each panel the ratio of the severe to the weak alleles decreases, indicating increased efficiency of purging the severe mutant allele, from the upper right (region I, the weak mutant has more negative epistasis) to the bottom left (region II, the severe mutant has more negative epistasis) part of the panel. The arrows A and B are discussed in the text.

A

Genotype	Frequency (Before selection, Generation T)	Relative fitness	Frequency (After selection, Generation T+1)
$\begin{bmatrix} A^S X^M & A^S X^{WT} \\ A^D X^M & A^D X^{WT} \\ A^{WT} X^M & A^{WT} X^{WT} \end{bmatrix}$	$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \end{bmatrix}$	$\begin{bmatrix} \omega_{11} = \omega_A^S \times \omega_X^M + \varepsilon & \omega_{12} = \omega_A^S \\ \omega_{21} = \omega_A^D \times \omega_X^M + \varepsilon & \omega_{22} = \omega_A^D \\ \omega_{31} = \omega_X^M & \omega_{32} = 1 \end{bmatrix}$	$\begin{bmatrix} \frac{p_{11}\omega_{11}}{\bar{\omega}} & \frac{p_{12}\omega_{12}}{\bar{\omega}} \\ \frac{p_{21}\omega_{21}}{\bar{\omega}} & \frac{p_{22}\omega_{22}}{\bar{\omega}} \\ \frac{p_{31}\omega_{31}}{\bar{\omega}} & \frac{p_{32}\omega_{32}}{\bar{\omega}} \end{bmatrix}$

where $\bar{\omega} = p_{11}\omega_{11} + p_{12}\omega_{12} + p_{21}\omega_{21} + p_{22}\omega_{22} + p_{31}\omega_{31} + p_{32}\omega_{32}$

B



Our results in Figure 2.2.3B depict the simulation results. The six panels in the figure represent the ratio of A^S to A^D alleles in the 50th, 100th, 150th, 200th, 250th, and

300th generations, respectively. Our simulations indicate that if the percentage of negative epistasis for the severe mutant is kept as a constant, as the percentage of negative epistasis for the weak mutation increases (as shown by the arrow A), the ratio of the severe to the weak allele frequency would increase. However, this ratio would decrease, indicating a faster removal of the severe mutants from the population, in another direction (as shown by the arrow B), i.e., the percentage of negative epistasis for the weak mutant is kept as a constant, but the percentage of negative epistasis for the severe mutant increases. Therefore, the distribution for the sign of epistasis among different alleles of the same gene observed in this study might represent an efficient way for eukaryotic organisms to purge deleterious mutations from populations.

2.3 Discussion

Our study represents a genome-wide theoretical survey for the dynamics of global epistatic effects under various mutant alleles of the same gene. We show that the epistatic profiling of a gene at the genome level is largely dependent on mutant types involved. Our results indicate that previous conclusions inferring epistatic relations among genes based on only one mutant type per gene can be greatly improved by using multiple mutant alleles. More importantly, our study shows that mutant alleles with severe defects have a higher chance to develop negative epistasis in eukaryotic organisms than in bacteria and archaea. It has been speculated that eukaryotic organisms might have more negative epistasis due to their increased complexity over prokaryotic organisms [114, 115]. Even if this hypothesis is true, however, our results for different mutant alleles of the same gene cannot be directly inferred from this complexity argument.

Even though the mechanism underlying our observation remains to be determined,

we argue that such distributions for negative epistasis among different alleles of the same genes have significant evolutionary consequences, as shown in our population genetics simulations (Figure 2.2.3). The origin and maintenance of sexual reproduction remains one of the central issues in evolutionary biology. Population genetics models have been proposed to explore the impact of epistasis on the maintenance of sexual reproduction [116, 117, 118, 119]. The mutational deterministic hypothesis posits that sex enhances the ability of natural selection to purge deleterious mutations by bringing them together into single genome through recombination [116]. This explanation requires the prevalence of negative epistasis at the genome level. Here we found that the mutations with larger deleterious defects within the same gene have a higher chance to develop negative epistasis in eukaryotic organisms than bacteria and archaea. The model we proposed in Figure 2.2.3, which is based on the population genetics theory from Kondrashov [116], indicates that such distribution of negative epistasis among different alleles of the same gene in eukaryotic organisms might lead to more efficient purging of deleterious mutations from populations, thus providing a previously unappreciated evolutionary advantage for sexual reproduction. We emphasize that these findings do not necessarily provide sufficient evidence to explain the cause for the emergence of sexual reproduction during evolution.

Although we found several unique characteristics regarding the global epistatic landscape of different mutant alleles in the same gene, three caveats need to be addressed. First, the FBA modeling used in this study, which has been successfully applied to various research problems [34, 93, 94, 95, 96, 97], includes only metabolic genes in the simulation. However, results from our analysis on the experimentally defined epistatic relations among $\approx$ 0.2 million double mutants comprising $\approx$ 4,000 *S. cerevisiae* genes, which nearly represent all functional categories in the budding yeast, confirmed our major FBA modeling predictions.

Second, even though FBA is one of the most comprehensive computational tools for simulating epistatic interactions among genes, there are still many aspects that can be improved to aid in capturing the full set of empirical genetic interactions [53]. For example, rules for transcriptional regulation and physical interactions can be integrated into the current FBA framework to improve its accuracy [120]. In addition, mapping between individual alleles and metabolic flux reduction is a complex process and difficult to measure experimentally [121]. It is noteworthy that in our simulations we have uniformly evaluated fitness consequence based on the percentage of WT flux attainable in a specific background. Depending on the regulation dynamics of individual genes, such uniform sampling may be unlikely to correspond to random sampling of mutant alleles. For instance, a mutation that limits the availability of a ligand that activates an enzyme following a Hill equation with early saturation may have a very high frequency of neutral or mildly deleterious mutations compared with a similar enzyme with late saturation. Nevertheless, uniform sampling in our study is still useful in illustrating the main evolutionary ideas presented here, which all have to do with relative severity of mutations rather than their absolute fitness.

Third, measuring the presence of epistasis is subject to a choice of threshold. Does the flux smoothly influence epistasis, or can epistasis abruptly change or become zero? We have seen evidence of both trends in our simulations. Though there are many different trends in the magnitude of epistasis that we are currently investigating, we present two cases to explore this issue (Figures A.6 and A.7 and Datasets S3 and S4). However, based on Figure A.1 and A.5, we have confirmed that our major results are robust to a variety of epistasis thresholds. As a result, although the choice of thresholds is a common problem for research on epistasis, we are still confident that our conclusion is unlikely to be significantly influenced by this factor. With these limitations in mind, our observations identified several important features for the epistasis among genes, and

call on future experimental and theoretical efforts to revisit genetics and evolutionary theories that can integrate epistatic dynamics among genes in biological systems.

2.4 Methods

2.4.1 Experimental Dataset

The experimental data were extracted from a global survey for the epistatic interactions among genes in *S. cerevisiae* [79]. In this original SGA study, the authors screened 1,712 *S. cerevisiae* query gene mutants against 3,885 array gene mutants to generate a total of more than 5 million gene mutant pairs spanning all biological processes. In each gene mutant pair, the epistasis value is calculated based on the equation $\epsilon = W_{xy} - W_x W_y$, in which W_{xy} is the fitness of an organism with two mutations in genes X and Y, and W_x or W_y refers to the fitness of the organism with mutation only at gene X or Y, respectively. In addition, a statistical confidence measure (p-value) was assigned to each interaction based on the observed variation of each double mutant across four experimental replicates and estimates of the background error distributions for the corresponding query and array mutants. Finally, a defined confidence threshold ($|\epsilon| \geq 0.01$, $P < 0.05$) was applied to generate epistatic interactions [79].

2.4.2 Flux Balance Analysis

FBA frames the stoichiometric equations that describe the biological reactions of a system as the following matrix equations, which is possible because stoichiometric equations are linear [8, 91, 92].

$$\begin{aligned} \text{maximize} \quad & \mathbf{c}^T\mathbf{v} \\ \text{subject to} \quad & \mathbf{S}\mathbf{v} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{0} \\ & \mathbf{v}_{lb} \preceq \mathbf{v} \preceq \mathbf{v}_{ub} \end{aligned} \tag{2.1}$$

The vector of concentration change over time ($\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$) is found by multiplying the stoichiometric matrix $\mathbf{S}$ by a flux vector $\mathbf{v}$. $\mathbf{S}$ has columns corresponding to each reaction in the system, and rows corresponding to metabolites. Typically, one or more enzymes correspond to each reaction, which allows us to see how a genetic perturbation, such as a knockout, may affect the system. The vector $\mathbf{v}$ consists of reaction fluxes and is subject to upper and lower bounds $\mathbf{v}_{ub} = (u_1, u_2 \ldots, u_n)^T$ and $\mathbf{v}_{lb} = (l_1, l_2 \ldots, l_n)^T$. If we want to simulate the knockout or knockdown of an enzyme, the fluxes corresponding to that enzyme can be constrained to be zero or lower than WT, respectively. It is assumed that the change in concentration over time is at steady state, therefore $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{0}$ in the FBA simulation [8].

The linear objective is written in terms of the v_i with weight coefficients c_i. Modified versions of COBRA and COBRA2 scripts, popular FBA software packages written for MATLAB, were used to implement our simulation framework [92]. The method for calculating a realistic WT flux for a given environment and organism model is taken from Smallbone and Simeonidis 8. This method, termed geometric FBA, attempts to choose a flux vector that is close to the average of all optimal flux vectors. The geometric FBA solution is also a minimal $\boldsymbol{L}^1$-norm solution, which has been previously heralded as a good choice because it minimizes the total amount of flux needed to achieve the objective, based on the fact that cells would avoid having much unnecessary flux and wasted energy [8]. A minimal $\boldsymbol{L}^1$-norm solution is advantageous in this study because restricting fluxes for mutants based on unnecessarily large WT fluxes may not constrain the system. Finally, the minimal $\boldsymbol{L}^1$-norm solution avoids the problem of having futile

cycles, which are thermodynamically infeasible [8].

Mutations of genes are simulated by the use of gene-reaction mapping and flux constraints. Enzymes may be involved in multiple reactions (i.e., pleiotropy). Although we often have Boolean rules describing the relationship between genes in an enzyme complex, it is currently extremely difficult to ascertain the exact contribution of each enzyme to each reaction [121]. Choosing the simplest unbiased approach, we used gene-reaction mapping and uniformly constrained the flux through each reaction associated to the gene being mutated. With one notable exception [34], most research relating to simulation of mutations with FBA has focused on null mutants [23, 93, 94, 95, 96, 97, 98]. Our simulation approach, though simplifying the actual dynamics that result in decreased fluxes in vivo, allows us to see behavior that was not previously possible. To be consistent, we used the same equation and threshold ($|\epsilon| \geq 0.01$) to calculate epistasis for FBA results as we did for the experimental data.

2.4.3 Population Genetics Model

A flowchart in Figure A.8 provides more illustration of the simulation procedure. We constructed a genetic system with a query gene A, which contains three different alleles (A^S: severe mutant; A^D: weakly deleterious mutants; and A^{WT}: WT) and a gene X that has two different alleles (X^M: mutant and X^{WT}: WT). The table in Figure 2.2.3A explains how genotype frequencies could be calculated from generation T to generation $T + 1$ under natural selection. In the figure, ρ and ω represent allele frequency and fitness, respectively. The average fitness in generation T could be calculated [122]. We simulated the ratio of allele frequency for the severe (A^S) to the weak (A^D) mutant alleles of the A gene under all possible combinations of the percentages of negative epistasis

for these two alleles, as shown on the *x* and *y* axis of Figure 2.2.3B. For each possible combination in each generation (a specific location on each panel of Figure 2.2.3B), the following two-step procedure was repeated 1,000 times. First, the epistatic relations (negative, positive, and no epistasis) between the mutant alleles of the genes A and X were randomly determined as the following: either A allele is assumed to have 10% possibility of having epistasis (either positive or negative) with the allele X^M [79]; when A and X alleles do have epistasis, the likelihoods for the epistasis being negative (and the remaining epistases are positive) are assigned independently for A^S and A^D alleles according to their location on Figure 2.2.3B. Second, the fitness of each genotype was calculated, which was then used to infer the genotype frequencies in the next generation according to Figure 2.2.3A. The average genotype frequencies among 1,000 randomizations were then recorded for simulations in the next generation. The ratio of allele frequency for the severe to the weak mutant alleles of the A gene in each generation was calculated based on genotype frequencies in that generation.

To make the simulation simple, the initial allele frequencies for the severe, weak, and WT alleles of the A gene were assumed to be equal (one-third), and the initial allele frequencies for the mutant and WT of the X gene were also assumed to be equal (one-half). The fitness was assumed to be 1, 0.99, and 0.98 for the WT, weak, and severe mutant alleles of gene A, respectively, and 1 and 0.99 for the WT and the mutant alleles of gene X, respectively. The positive and negative epistasis values between A and X gene mutants were assumed to be 0.01 and 0.01, respectively. A variety of fitness differences between the severe and weak alleles and epistasis values have also been used in the simulations, and the trend remains the same.

2.5 Acknowledgments

We thank Dr. Ricardo Azevedo for his insights and critical comments on the paper; Dr. Huifeng Jiang and Mr. Kaixiong Ye for discussion; and the editor and two anonymous reviewers for constructive comments. This work was supported by a startup fund from Cornell University, National Science Foundation Grant DEB-0949556, and National Institutes of Health Grant 1R01AI085286 (to Z.G.).

CHAPTER 3

DYNAMIC EPISTASIS UNDER VARYING ENVIRONMENTAL PERTURBATIONS

Epistasis describes the phenomenon that mutations at different loci do not have independent effects with regard to certain phenotypes. Understanding the global epistatic landscape is vital for many genetic and evolutionary theories. Current knowledge for epistatic dynamics under multiple conditions is limited by the technological difficulties in experimentally screening epistatic relations among genes. We explored this issue by applying flux balance analysis to simulate epistatic landscapes under various environmental perturbations. Specifically, we looked at gene-gene epistatic interactions, where the mutations were assumed to occur in different genes. We predicted that epistasis tends to become more positive from glucose-abundant to nutrient-limiting conditions, indicating that selection might be less effective in removing deleterious mutations in the latter. We also observed a stable core of epistatic interactions in all tested conditions, as well as many epistatic interactions unique to each condition. Interestingly, genes in the stable epistatic interaction network are directly linked to most other genes whereas genes with condition-specific epistasis form a scale-free network. Furthermore, genes with stable epistasis tend to have similar evolutionary rates, whereas this co-evolving relationship does not hold for genes with condition-specific epistasis. Our findings provide a novel genome-wide picture about epistatic dynamics under environmental perturbations.

3.1 Author Summary

Epistasis, often referred to as genetic interactions, occur when mutational effects of genes depend on each other. Aside from often times complicating the way in which the

phenotype of an organism relates to its genotype, epistatic interactions (or epistases) are essential to several important theories in biology, especially in evolution. Due to the difficulty in experimentally assessing epistasis across an entire genome, we employed mathematical modeling of the metabolic network of bakers yeast to comprehensively simulate genetic interactions for virtually all known metabolic genes in the organism. We performed comprehensive simulations in 17 different environments, which differ by their nutrients. We characterized a trend that occurs in genetic interactions when yeast is transferred from a glucose-abundant environment to other environments. We also found that both the set of genetic interactions present in all conditions and the set of interactions present in a single environment are fairly large sets with highly different connectivity. Furthermore, the set present in all conditions tends to consist of gene pairs with similar evolutionary rates.

3.2 Introduction

Epistasis refers to the phenomenon wherein mutations of two genes can modify each others phenotypic outcomes. It can be positive (alleviating), or negative (aggravating), when a combination of deleterious mutations shows a fitness value that is higher, or lower, than expectation, respectively. For example, a mutation that hampers a pathway's function may allow for other mutations in the same pathway without a fitness consequence, resulting in positive epistasis. Conversely, genes or pathways with redundant functions can give rise to negative epistasis. It is well established that epistasis is important for the evolution of sex [107, 117, 123], speciation [124], mutational load [125], ploidy [126], genetic architecture of growth traits [127], genetic drift [128], genomic complexity [115], and drug resistance [129]. As biological systems in nature have to face multiple genetic and environmental perturbations, understanding the global land-

scape and dynamics of epistasis under these perturbations remains an important issue in the evolutionary field. In an earlier study, we addressed genome-wide epistasis dynamics under various genetic perturbations [9]. In this study, we will investigate the impact of environmental perturbations on global epistasis dynamics.

How epistatic interactions among genes change in different environmental conditions has been intensively studied in various model organisms, including *E. coli* [130, 131, 132], *S. cerevisiae* [133, 134, 135], *C. elegans* [136, 137] and *D. melanogaster* [138, 139, 140]. The results of these studies, however, are very controversial. While some studies observed increasing positive epistasis under harsh conditions [131, 135, 138], others have opposite findings [132, 133, 134, 136, 137, 139, 140, 141]. Even within the same species, different experimental studies might have conflicting conclusions (e.g. Kishony and Leibler 131, Cooper et al. 132). One possible reason for the above controversy could have originated from the fact that most studies only looked at the epistasis dynamics based on a small number of genes, where the properties cannot be generalized to the entire organism.

The main obstacle to exploring global epistatic dynamics under a variety of environments is the difficulty of applying high-throughput experimental platforms. To explore epistasis on a genomic scale, a number of technologies have been developed to systematically map genetic interaction networks, such as synthetic genetic array (SGA) [79, 80], diploid-based synthetic lethality analysis with microarrays (dSLAM) [81, 82], synthetic dosage-suppression and lethality screen [83, 84, 85] and epistatic miniarray profiles (EMAP) [86, 87, 88]. A key issue for all these experimental studies is that these epistatic networks have been constructed only under normal laboratory conditions. However, cells are constantly bombarded by various external environmental stresses. Epistasis dynamics under these perturbations cannot be predicted based on normal lab-

oratory conditions. Few studies have constructed epistatic networks under multiple environmental perturbations. A recent study that has only constructed epistatic networks for a group of genes with specific functions under one normal and one harsh condition already requires a large amount of effort [142]. Consequently, genome-scale epistasis landscapes under a variety of environmental perturbations remain largely uncharacterized.

Here we explored this issue by using Flux Balance Analysis (FBA) to simulate epistasis dynamics among genes under multiple environmental perturbations. FBA involves optimization of an objective function, commonly growth maximization in microbes, subject to the reactions and constraints of a metabolic network, which can provide reliable predictions [8, 22, 24, 92, 98, 101]. Using this platform, a previous study has investigated synthetic lethal interactions (one type of negative epistasis) under multiple environmental perturbations and showed the plasticity of epistatic interactions in the metabolic networks [95]. Here we examined both positive and negative epistasis using FBA, and were able to show that at the genome scale epistatic interactions tend to become more positive in nutrient-limiting conditions relative to abundant-glucose media. In addition, while a large proportion of epistatic interactions can be rewired dynamically under varying environments, there is a core of epistatic interactions that are stable across all tested environments. We also discovered different network and evolutionary properties for genes with stable and dynamic epistatic interactions. Implications of our findings were discussed.

3.3 Results

3.3.1 FBA modeling and simulated growth conditions

We applied the yeast *S. cerevisiae* metabolic reconstruction iMM904 [91] to examine the dynamics of epistasis under various environmental perturbations. The reconstruction has 904 metabolic genes that are associated with 1,412 metabolic reactions. We conducted FBA simulations under an abundant-glucose condition and 16 nutrient-limiting conditions with the following environmental perturbations. In 15 of these perturbations, the carbon source (abundant glucose) was replaced by one of the following: acetaldehyde, acetate, adenosine 3',5'-bisphosphate, adenosyl methionine, adenosine, alanine, allantoin, arginine, ethanol, glutamate, glutamine, glycerol, low glucose, trehalose, and xanthosine, respectively. These conditions represent a wide variety of nutrient and energy sources: nucleosides, amino acids, sugars, alcohols, etc. Additionally, we looked at abundant glucose under limited phosphorus availability.

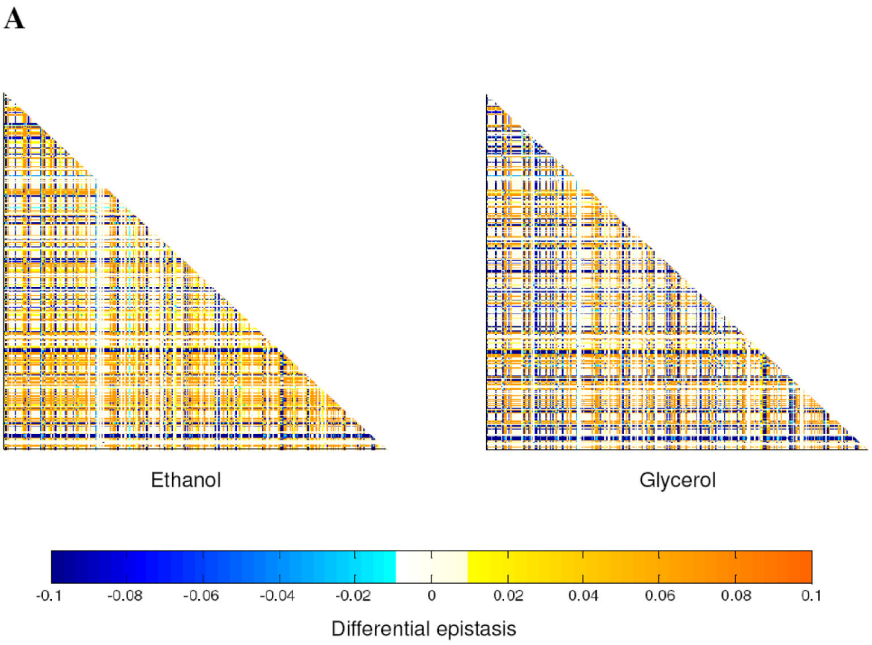
To ensure that all these environmental conditions have the same growth rates in the following analyses, we restricted the carbon source or phosphorous uptake levels for each of the 16 environmental perturbations such that only 20% of the high-glucose growth rate was attained. This was chosen because it has been shown that metabolism was directly linked to growth and similar growth rates often induced similar metabolic pathways [143]. It is therefore important to use a fixed growth rate among different conditions to control for the relationship between growth rates and the overall metabolic activity so as not to induce a growth-rate specific effect. The 20% high-glucose level was chosen because some media types do not support high growth rates, regardless of the abundance of the nutrient source. In order to estimate epistasis between genes, we created a mutation for each gene in each condition that restricted the flux to be

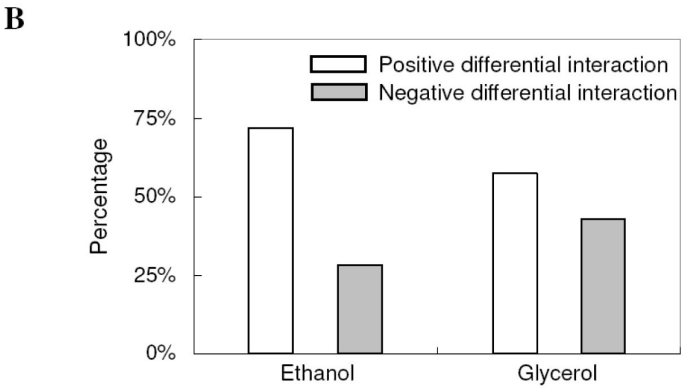
50% of the wild-type flux found by geometric FBA for all reactions associated with the mutant gene [9]. Epistatic relations between any two genes were calculated under each condition. We also tested our core findings allowing maximum growth in each condition (Table S1) and the general trends in our results remained similar, as described in the following.

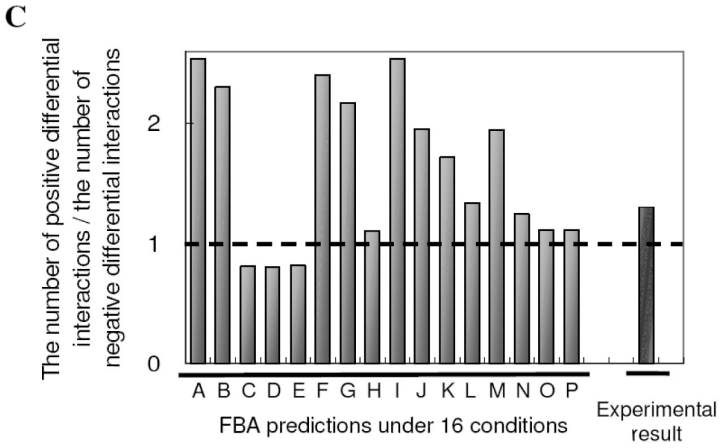
3.3.2 More positive differential epistases from rich media to nutrient-limiting conditions

To directly address how the sign and magnitude of epistases change under nutrient-limiting conditions, we calculated differential epistasis ($d\epsilon$), which is defined as the epistatic changes from abundant-glucose media to the nutrient-limiting condition for each gene pair in each growth condition. A gene pair with positive (or negative) differential interaction under an environmental perturbation is defined as these two genes having increasing (or decreasing) epistasis values from the abundant-glucose media to that condition. Figure 3.3.2A depicts the distribution of differential epistases in two growth conditions (ethanol and glycerol) as an example. Only genes with $|d\epsilon| \geq 0.01$ in at least one of the two conditions are included in this figure. As quantified in Table S2, there are 6.1% and 5.5% of total gene pairs with $|d\epsilon| \geq 0.01$ from abundant-glucose media to ethanol and glycerol growth conditions, respectively. Among them, a large number of gene pairs even change their sign of epistasis (Table S2). Simulations in other conditions show similar effects (Table S2), indicating that epistatic relationships among genes can be very dynamic between abundant-glucose media and nutrient-limiting conditions.

Figure 3.1: More positive differential epistases under environmental perturbations. (**A**) Heat maps describe the global dynamics of differential epistasis from abundant-glucose medium to ethanol (left panel) and glycerol (right panel) conditions. Only gene pairs with $|d\epsilon| \geq 0.01$ in either condition are included in the figure. Different colors represent differential epistasis values as indicated by the color bar at the bottom. The differential epistasis values are assigned to be 0.1 (or -0.1) in the heat-maps when it is greater than 0.1 (or less than -0.1). It is noteworthy to point out that the epistasis patterns are indeed very different between the two conditions (Figure 3.2A). (**B**) Percentage of positive and negative differential epistases under ethanol and glycerol conditions. (**C**) Ratio of positive to negative differential epistases in each simulated condition. The result from a high-throughput experiment is also shown. The letters A-P represent acetaldehyde, acetate, adenosine 3',5'-bisphosphate, adenosyl methionine, adenosine, alanine, allantoin, arginine, ethanol, glutamate, glutamine, glycerol, low glucose, phosphate, trehalose, and xanthosine, respectively.


A
Ethanol
Glycerol
-0.1
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
0.1
Differential epistasis


B
100%
75%
50%
25%
0%
Percentage
Positive differential interaction
Negative differential interaction
Ethanol
Glycerol


C
The number of positive differential interactions / the number of negative differential interactions
0
1
2
A B C D E F G H I J K L M N O P
FBA predictions under 16 conditions
Experimental result

We further investigated the sign of differential epistasis from abundant-glucose to nutrient-limiting conditions. As shown in Figure 3.3.2A, we observed more yellow dots (positive differential epistasis) than blue dots (negative differential epistasis) in both panels. Indeed, as quantified in Figure 3.3.2B, 72% and 57% of differential epistases are positive in ethanol and glycerol conditions, respectively. We further explored all 16 nutrient-limiting conditions and the results are shown in Figure 3.3.2C. In most of our simulated conditions (13/16), there are significantly more positive differential epistases than negative differential epistases (Binomial test, $P < 10\text{-}5$ for each of the 13 conditions), indicating that epistasis tends to become more positive in nutrient-limiting conditions. This conclusion does not depend on the criteria we used to define differential epistasis (Figure B.1).

A recent high-throughput experiment measured epistatic relations between $\approx$ 80,000 gene pairs with and without perturbation by a DNA-damaging agent (methyl methanesulfonate, MMS). The study represents the most comprehensive experimental study so far to explore epistatic dynamics from a rich medium to a harsh condition [142]. Interestingly, the authors also found more positive differential epistases than negative differential epistases, which is consistent with our general observation (Figure 3.3.2C). We further allowed maximum growth in each condition and the general trends in our results remained similar (Figure B.1).

We found that differential epistasis had functional importance after performing both Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses to compare genes with positive and negative differential epistases through the glucose-abundant to ethanol transition. We chose the ethanol condition as an example because it is one of the most widely used conditions for the bakers yeast. Interestingly, we observed that 38 GO terms and 8 KEGG pathways are enriched for

positive differential epistasis, while 18 GO terms and 1 KEGG pathways are enriched for negative differential epistasis (Table S3). More importantly, we found positive and negative differential epistases uniquely contribute to different aspects of ethanol and energy metabolism. For example, positive differential epistasis is enriched in monohydric alcohol metabolic processes, oxidoreductase activity acting on aldehyde group donors, the TCA cycle, and pyruvate metabolism, while negative differential epistasis is enriched in ethanol metabolic processes and various amino acid terms and pathways, indicating the functional importance of differential epistasis (Table S3). This is consistent with experimental results that show differential epistatic interactions, rather than static epistatic interactions, are functionally related to the response of interest [142].

Several system properties were found to correlate with the ratio of the number of positive to negative differential epistases (Table S4). A strong correlation exists between the number of essential genes in a given condition and the ratio of positive to negative differential epistasis on transition from high glucose to that condition (ρ = 0.9056, P = 1.3905e-006), which was a better predictor than the number of non-zero fluxes in the wild-type vector for that environment (ρ = 0.7131, P = 0.0019). An even stronger predictor for positive differential epistasis was the mean relative fitness of single mutants in the new environment (ρ = −0.9941, P = 6.4340e-015); this anticorrelation suggests that a propensity for a lower single mutant fitness can cause a shift towards positive epistasis.

3.3.3 Dynamic epistasis between nutrient-limiting conditions

Figure 3.3.2 explored the epistasis dynamics from abundant-glucose media to nutrient-limiting conditions. As biological systems in nature constantly face changing envi-

ronmental perturbations, it is interesting to investigate the epistasis dynamics among nutrient-limiting conditions. To achieve this aim, we first explored the epistatic relationship between the same gene pairs in the two environmental perturbations based on growth in ethanol and glycerol. Figure 3.2A lists the number of gene pairs that have various epistatic relationships. It is noteworthy to point out that, consistent with previous published results there are significantly more positive epistases between genes than negative ones in either condition [34].

A

		Ethanol		
		Positive Epistasis	Negative Epistasis	No Epistasis
Glycerol	Positive Epistasis	24,567	159	1,469
	Negative Epistasis	752	182	432
	No Epistasis	2,647	67	378,785

B

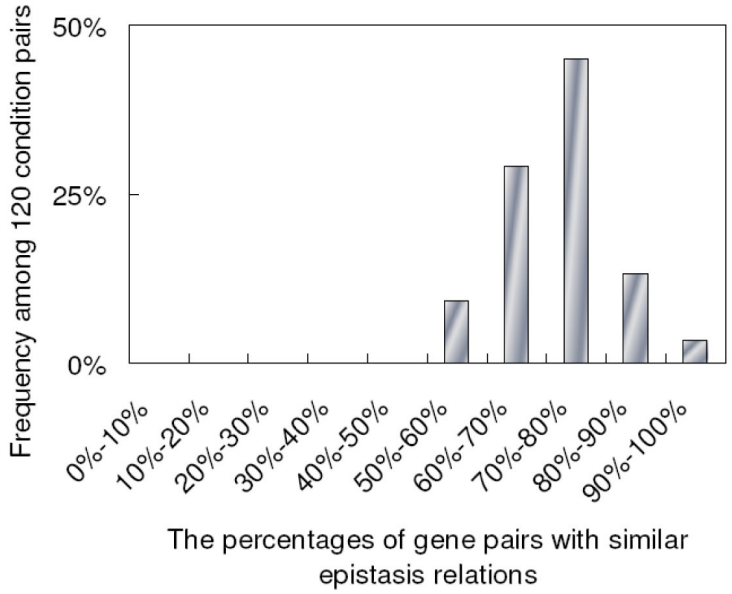


Figure 3.2: Epistasis dynamics between environmental perturbations. (**A**) Number of gene pairs with various epistatic relationships between ethanol and glycerol growth conditions. (**B**) The distribution for the percentages of gene pairs with similar epistasis relation between any 2 of 16 conditions. The frequency is derived from the 120 pairs of environmental conditions simulated in this study.

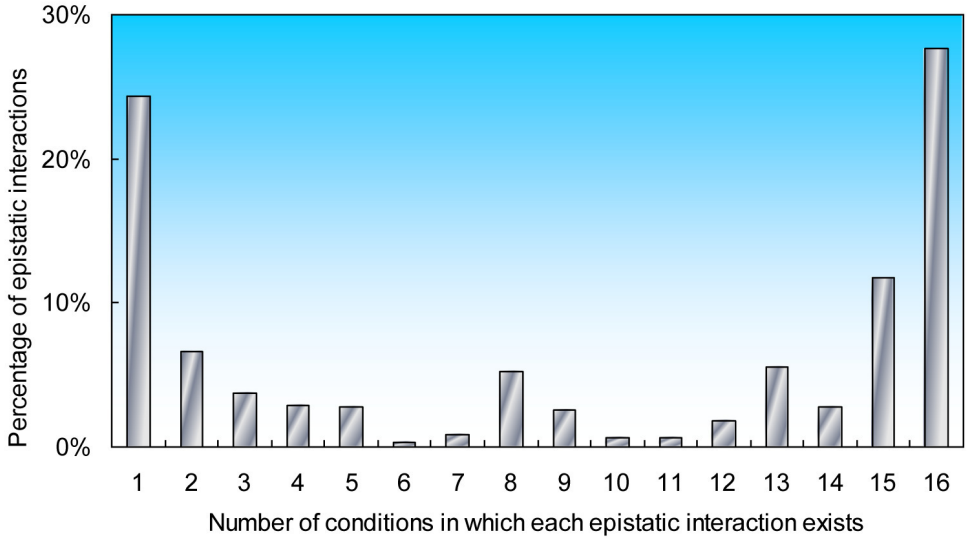
If two genes have the same sign of epistasis and $|\epsilon| \geq 0.01$ in both conditions, they are defined as having similar epistatic relationship in these two conditions. To quantify epistatic dynamics between ethanol and glycerol growth conditions, we defined the per-

centage of gene pairs with similar epistatic relations to be the number of gene pairs with similar epistasis relations shared in these two conditions (overlap) divided by the number of gene pairs with epistasis in either condition (union). Our results show that 79% of gene pairs have similar epistasis relations between these two conditions. Figure 3.2B shows the distribution for the percentages of gene pairs with similar epistasis relations between any 2 of 16 conditions, demonstrating a variable degree of epistatic similarity between any two conditions. This conclusion still holds when we used different criteria to define epistatic relationships between genes (Figure B.3).

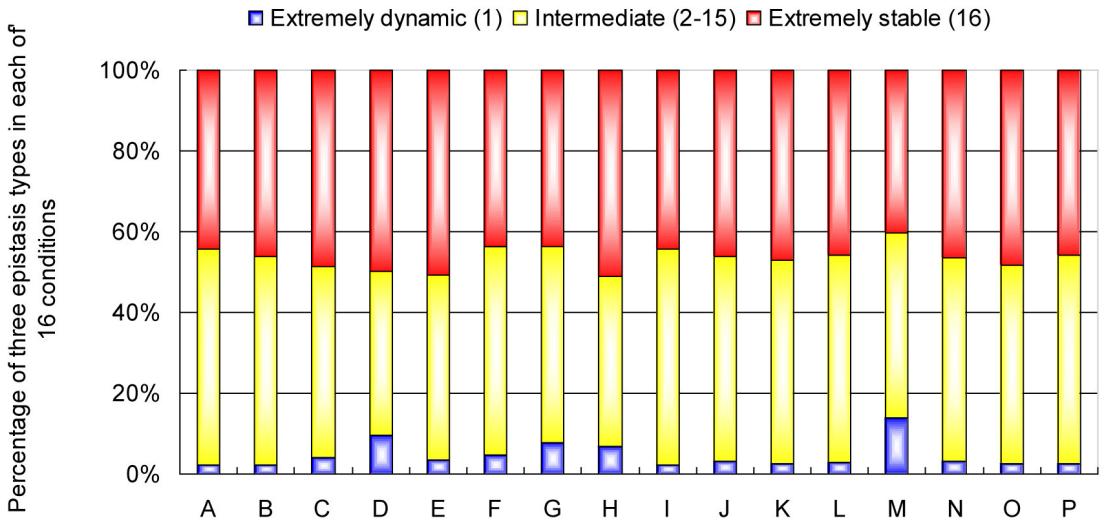
To understand the global distribution of all epistatic relations, we considered 16 conditions together and calculated the fraction of epistatic interactions existing in 1, 2, 3, ..., 15, and 16 conditions, respectively. As shown in Figure 3.3.3A, we found that there is a U frequency distribution for the number of growth conditions in which a specific epistatic interaction is observed. This means that approximately 52% of these interactions are either condition-specific (24%; termed dynamic) or predicted to exist in all conditions (28%; termed stable), and $\approx$ 48% is intermediate (exists in multiple but not all 16 conditions). An analogous result was obtained previously, but only for synthetic lethal interactions [95]. We also changed the growth assumption and allowed maximum growth in each condition and reanalyzed the global distribution of all epistatic relations. The U frequency distribution for the number of growth conditions in which a specific epistatic interaction is observed remained similar (Figure B.4). Based on the result in Figure 3.3.3A, we further calculated the ratio of these three types of epistatic relations in each of the 16 environmental perturbations. As shown in Figure 3.3.3B, we found that in each environment, $\approx$ 40-60% of epistatic interactions are stable ones and that each specific environmental condition also has many private epistases among genes.

Figure 3.3: The global distribution of epistatic relations under simulated conditions. (**A**) Distribution for the number of conditions in which each epistatic interaction exists. Note that ≈ 28% of epistatic relations are extremely stable (the very right bar) and ≈ 24% are extremely dynamic (the very left bar). (**B**) Fraction of three types of epistatic relations in each of the 16 environmental perturbations, as indicated by the color bar to the right. The numbers in the brackets represent the number of conditions in which each epistatic interaction exists, as indicated in (**A**). The letters A-P represent the simulated conditions as indicated in Figure 1.

A


30%
20%
10%
0%
Percentage of epistatic interactions
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Number of conditions in which each epistatic interaction exists

B


Extremely dynamic (1)
Intermediate (2-15)
Extremely stable (16)
Percentage of three epistasis types in each of 16 conditions
100%
80%
60%
40%
20%
0%
A B C D E F G H I J K L M N O P

3.3.4 Different network properties for stable and dynamic epistasis

Analysis on network properties can reveal various organization principles (e.g. frequency of occurrence, centrality) for epistasis networks [79, 80] and therefore provide valuable information to further distinguish stable and dynamic epistasis. To achieve this aim, we compared networks formed by extremely stable and extremely dynamic epistasis among genes and asked whether they have distinct network properties. The degree distributions for both types of epistasis are shown in Figure 3.4A. Interestingly, extremely stable epistatic interactions form an exponential network architecture, which is homogeneous, meaning that most nodes have a very similar number of links (Figure 3.4A, left panel). In contrast, the extremely dynamic epistatic interactions give rise to a scale-free network topology, which is heterogeneous, meaning that the majority of nodes have few links but a small number of hubs have a large number of links (Figure 3.4A, right panel).

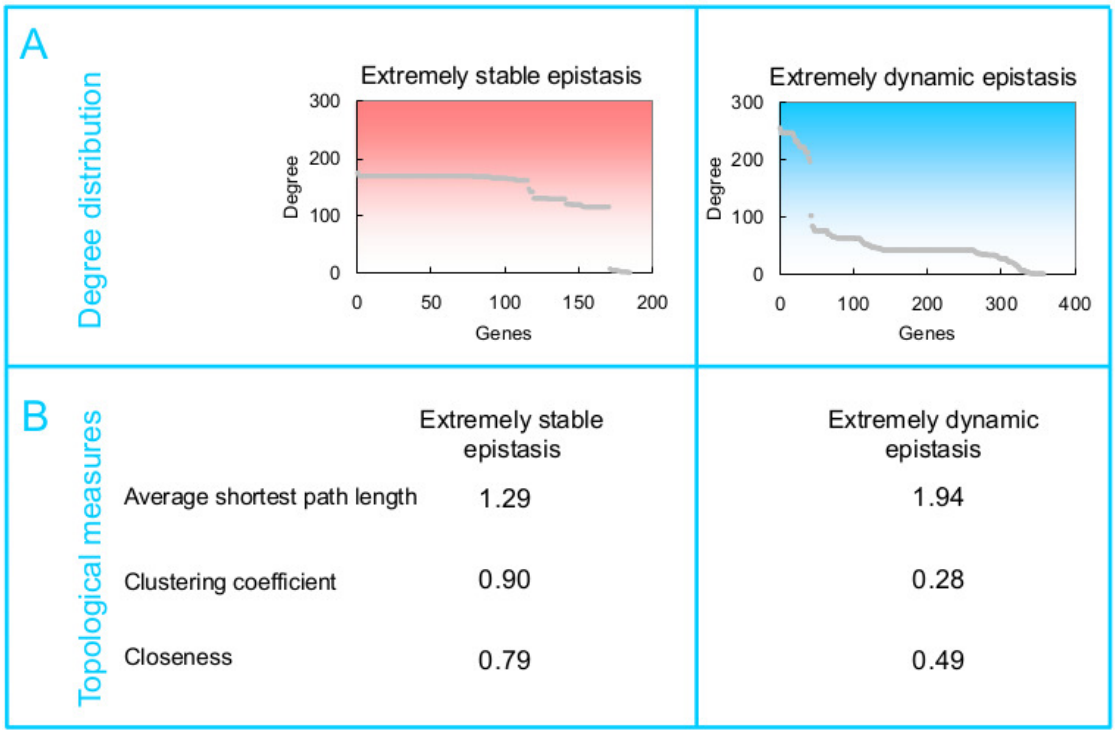


Figure 3.4: Network properties for the extremely stable and extremely dynamic epistatic interactions. (**A**) Degree distribution for genes in two epistatic interaction networks. The networks have nodes that correspond to genes and edges that correspond to epistatic interactions. (**B**) Three network parameters (the definition of which are shown in Methods) for two epistatic interaction networks.

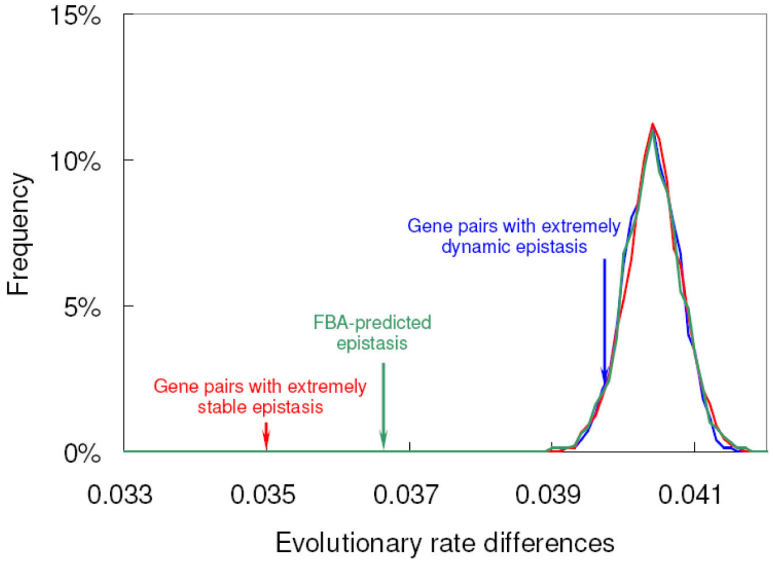
In addition, we calculated three network parameters to compare these two types of epistatic interactions. We found that the network formed by extremely stable epistases has a smaller shortest path length, a larger clustering coefficient and larger closeness than the network formed by extremely dynamic epistases (Figure 3.4B). These results are consistent with the scenario that genes with extremely stable epistasis are directly linked to most other genes and form an exponential network topology, while genes with extremely dynamic epistasis form a scale-free network. Our results also show that the network induced by intermediate epistases have intermediate values of these parameters compared to that of extremely stable and extremely dynamic epistasis networks (Tables S5 and S6).

3.3.5 Co-evolution of genes with epistatic interaction

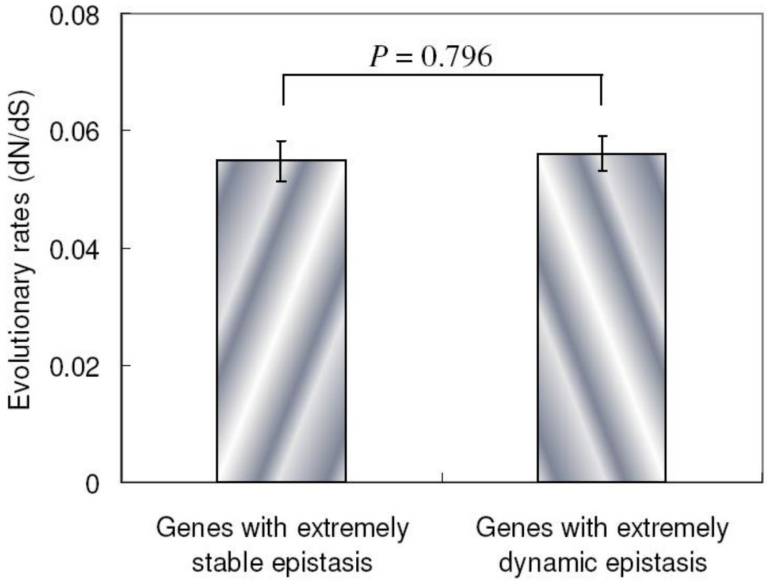
Gene pairs with epistasis identified in real experiments usually show similar evolutionary rates. To investigate whether two genes with predicted epistasis also tend to co-evolve, we calculated the evolutionary rate differences between two genes with epistasis from FBA modeling (Figure 3.3.5A). Evolutionary rates ($^{d_N}/_{d_S}$) based on orthologous gene sets from four yeast species of the genus Saccharomyces were downloaded from a commonly used reference dataset [144]. Simulations based on the same number of gene pairs with FBA-predicted epistasis were conducted to estimate the evolutionary rate differences for any two randomly selected genes. As shown in Figure 3.3.5A, the gene pairs with FBA-predicted epistatic interactions tend to have more similar evolutionary rates than random expectation (P < 10-4).

Figure 3.5: Co-evolution between genes with epistasis. (**A**) Average evolutionary rate differences between gene pairs with FBA-predicted epistasis (green), extremely dynamic epistasis (blue) and extremely stable epistasis (red) are highlighted by three arrows, respectively. The random simulations with the same number of gene pairs as each of the three groups were repeated 10,000 times and the frequency distributions are shown (marked by the same colors as the corresponding arrows, respectively). (**B**) The evolutionary rates for genes that are involved in extremely stable and extremely dynamic epistasis, respectively. The error bars represent standard errors.

A


15%
10%
5%
0%
Frequency
Gene pairs with extremely dynamic epistasis
FBA-predicted epistasis
Gene pairs with extremely stable epistasis
0.033
0.035
0.037
0.039
0.041
Evolutionary rate differences

B


0.08
0.06
0.04
0.02
0
Evolutionary rates (dN/dS)
P = 0.796
Genes with extremely stable epistasis
Genes with extremely dynamic epistasis

In Figure 3.4 we observed unique network properties for extremely stable and extremely dynamic epistatic interactions. We further investigate the co-evolution between genes with these two types of epistatic relationships. As shown in Figure 3.3.5A, genes with extremely stable epistasis tend to co-evolve (P < 10-4), while the difference between genes with extremely dynamic epistasis and random expectation becomes much smaller (P = 0.06). The evolutionary rate difference between gene pairs with extremely stable and extremely dynamic epistasis is also significant (*t*-test, P = 8e-6). This difference is not caused by genes that are involved in extremely stable or extremely dynamic epistasis, because these two groups of genes do not have significantly different evolutionary rates (*t*-test, P = 0.796, Figure 3.3.5B).

3.4 Discussion

3.4.1 Natural selection in nutrient-limiting conditions

Whether a genetic mutation has a fitness consequence depends on other sites, a phenomenon called epistasis (see Lehner 145 for a recent review on molecular mechanisms). Positive epistasis alleviates the total harm when multiple deleterious mutations combine together and thus reduces the effectiveness of natural selection in removing these deleterious mutations, whereas negative epistasis plays the opposite role by increasing the efficiency of purging deleterious mutations by natural selection. Results from this study present an initial glimpse over environment-induced epistasis dynamics at the genome scale. Using differential epistasis from abundant-glucose to nutrient-limiting conditions, our results show that epistasis between specific genes can become more positive or more negative in nutrient-limiting conditions, which is consistent with

previous findings in small scale studies [130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141]. However, we showed that, at the genome scale, epistasis is more positive in nutrient-limiting conditions. Interestingly, our simulation results are consistent with a recent genome-wide study between laboratory and harsh growth conditions [142]. How epistasis affects selection in harsh conditions has been controversial [146]. Our results provide the genome-wide evidence arguing that selection might be less effective in removing deleterious mutations in harsh conditions, which could be one of the underlying reasons for a recent observation that stimulation of a stress response can reduce mutation penetrance in *Caenorhabditis elegans* [147].

3.4.2 Network properties and evolutionary patterns for stable and dynamic epistasis

Our results indicate that epistasis could be extremely stable or dynamic among various environmental perturbations, which is consistent with a previous FBA study investigating synthetic lethal relations among non-essential genes [95]. The inclusion of essential genes in our study allows for investigation on many important metabolic pathways that were not previously analyzed. Nevertheless, the distribution of epistasis among multiple environments (Figure 3.3.3A) remains largely unchanged from the previous study [95] even when essential genes are included.

We also found that stable and dynamic epistatic relationships show totally different network properties and evolutionary patterns, which might provide new biological and evolutionary insights. The gene pairs with stable epistases tend to co-evolve with each other. In addition, from the biological pathway perspective, the smaller shortest path length and larger closeness values in the stable epistasis network both imply that genes

with stable epistases tend to be functionally associated with a large number of neighbors to form a condensed functional network, different from genes in the dynamic epistasis network that are loosely connected. Furthermore, the large clustering coefficient in the stable epistasis network also supports the idea that genes with stable epistasis interactions form a network core in the whole epistasis network. Combined with observations in Figure 3.3.5, this core module of epistasis in the metabolic network might represent stable functional associations between genes that are essential for important biological functions and evolutionary conserved even under different environmental perturbations. The lack of co-evolution pattern and scale-free network properties for the dynamic epistasis network, however, might represent unstable functional associations between genes, which may only be responsible for unique functions under specific conditions.

3.4.3 Implications and significance for exploring stable and dynamic epistasis

Our prediction about stable and dynamic epistasis could have important functional applications. A recent study showed that the synthetic lethal (negative epistasis) relationship between fumarate hydratase and haem oxygenase can be employed successfully to identify an in vitro drug target in hereditary leiomyomatosis and renal-cell cancer (HLRCC) cells [62]. Exploring both dynamic and stable epistasis could be useful in this context; stable epistatic interactions may be important for drug target detection in cancer or other pathogens, whereas it may sometimes be necessary to exploit dynamic epistatic relationships, possibly induced by treatment with an external perturbation.

Furthermore, rational evolutionary design techniques such as OptKnock [148] and OptGene [149] attempt to find which knockouts will enable a reaction of interest to

be coupled with growth (i.e. have positive epistasis with growth-associated genes in a specific environment). However, these techniques do not take into account epistasis dynamics across different environments. In this study, we have found that epistatic relations can be highly dynamic under various environmental perturbations, which raises the possibility to improve these techniques by considering epistasis dynamics in future studies. Research on using compensatory perturbations to reach desired network states is ongoing [150].

3.4.4 Caveats and future directions

Though we show several novel insights into how varying environments can influence epistasis, several caveats should be addressed. First, the FBA modeling used in this study, which was proven to have great predictive power and has been successfully employed in addressing numerous research problems [8, 24, 98], only includes metabolic genes. Second, even though FBA offers the most comprehensive simulation method for studying epistasis, there are many improvements that can be made in order to capture the empirically observed set of epistatic interactions [53]. For example, integrating transcriptional regulation and physical interactions into this framework could improve the current methods in predicting epistasis and other evolutionary processes [120]. Related to this point, FBA as used herein only considers the steady state and does not take into account any dynamics or initial conditions, and would necessarily miss any epistatic interactions that are due to dynamics in the system, such as changing concentrations; dynamic FBA (which is part of rFBA) might be a solution, but would likely require about a minimum of two orders of magnitude increase in computation time [20, 53]. Recent work on new objective functions targeting metabolite turnover rather than flux per se has also proven successful in recovering many epistases that were previously not

found with FBA [151].

Third, in order to understand the impact of environmental perturbations on epistasis, we used a reductive approach and only considered one mutation type per gene to simulate the global epistatic landscape in 16 environments. There are countless environments in nature. Furthermore, different mutations in the same gene and the interactions between genes and environment can likely have an even more complex impact on the epistasis dynamics. While it would be ideal to simulate a larger variety of environmental conditions for multiple mutations of the same gene, the computational cost is a limiting factor. Our previous study showed that different mutants of the same gene could have very dynamic epistatic interaction partners in a single environment [9]. In this study, we chose to use one mutation per gene as we are focusing on addressing how different environments could affect gene epistasis dynamics. Nevertheless, in order to see how sensitive our results were, we performed the analysis for our core results by simulating 16 environments using different growth assumption, where the organisms are allowed to have unrestricted uptake of the limiting nutrient to obtain the maximum growth in that condition. We found the major trends in our results are largely unchanged (Figures B.1 and B.1; Table S1).

Keeping these issues in mind, our analysis uncovered several prominent features of epistatic interactions under a variety of environmental perturbations, and call on future effort to confirm these simulation results using high-throughput experimental platforms. More importantly, the enrichment of stable and dynamic epistasis provides a new perspective to understand how biological systems may rewire epistasis in nature.

3.5 Methods

Scripts for generating and analyzing the data can be found in the source code repository located at `https://github.com/bbarker/COBRAscripts/`. Scripts and documentation specific to this paper are located in the subdirectory `MyProjects/EnvironmentalEpistasisFBA`.

3.5.1 Flux Balance Analysis

Flux Balance Analysis attempts to tackle issues inherent in other methods of metabolic modeling, such as the need to measure a large number of parameters, slow speed of simulation, and dependence on initial conditions [22, 152]. Other than needing a fairly complete understanding of the reactions present in an organism, the only measurements required to perform a genome-scale metabolic simulation are those for determining biomass constitution or a gene expression profile [24, 91]. Strictly speaking, FBA is a particular type of constraint based modeling (CBM). Constraint based modeling frames the stoichiometry that describe the reactions present in an organism as a matrix equation with indeterminates (reaction fluxes) subject to constraints [8, 91]. The optimization problem is described as follows:

$$\begin{aligned} \text{maximize} \quad & \mathbf{c}^T\mathbf{v} \\ \text{subject to} \quad & \mathbf{S}\mathbf{v} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{0} \\ & \mathbf{v}_{lb} \preceq \mathbf{v} \preceq \mathbf{v}_{ub} \end{aligned} \tag{3.1}$$

S is a matrix, in which rows and columns correspond to cellular metabolites and reactions in the reconstructed network respectively. **v** is the reaction flux with upper and

lower bounds $\mathbf{v}_{ub}$ and $\mathbf{v}_{lb}$ respectively. Multiplying the stoichiometric matrix **S** by the flux vector v equals the concentration change over time ($\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$). At steady state, the flux through each reaction is given by $\mathbf{Sv} = \mathbf{0}$. Further details on the underlying methods can be found in the literature [8, 9, 34].

The fluxes of mutations employed in this analysis were restricted to be 50% of the wild-type fluxes found for growth rate maximization by geometric FBA [34]. To find new conditions with a specified carbon source or other limiting nutrient that achieves 20% of the high-glucose growth rate, we can solve a linear program for the minimization of the limiting nutrient uptake while requiring the growth rate to be equal to 20% of the abundant-glucose growth-rate. For maximum growth rate conditions (Table S1, Figures B.1 and B.4), we allowed unrestricted uptake of the limiting nutrient to obtain the maximum growth in that condition, up to the point where it would reach the high-glucose growth rate. Mutations affecting protein complexes and pleiotropic genes are handled by uniform restriction across enzymes as described before [9].

3.5.2 Definition of epistasis

In each gene mutant pair, the epistasis value is calculated based on the equation: $\epsilon = W_{xy} - W_x W_y$, in which W_{xy} is the fitness of an organism with two mutations in genes X and Y, whereas W_x or W_y refers to the fitness of the organism with mutation only at gene X or Y respectively. Each fitness listed previously is calculated relative to the wild-type fitness. Absolute fitness values are determined by the value of the biomass maximization objective present in the model. Finally, a confidence threshold ($|\epsilon| \geq 0.01$) was applied to generate epistatic interactions [9, 34, 79]. We also conducted analyses based on a different threshold for epistasis and the general conclusions still hold in our analysis

(Figures B.1 and B.3).

3.5.3 Evolutionary rates and network parameters

Evolutionary rates of *S. cerevisiae* genes were downloaded from supplementary materials of Wall et al. 144, in which orthologs were defined by four complete genomes of *Saccharomyces* species (*Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae* and *Saccharomyces bayanus*) and evolutionary rates at synonymous and nonsynonymous sites were calculated based on a four-way yeast species alignment for *S. cerevisiae* genes by PAML. For the distributions in 3.3.5A, we randomly sampled gene pairs with the same number of gene pairs as in three epistasis networks (epistasis in all 16 conditions, extremely stable epistasis, and extremely dynamic epistasis, respectively), and calculated the average evolutionary rate differences between random gene pairs in each of these three sample sets. The simulations were repeated 10,000 times for each of the three groups, which are color coded to correspond to the epistasis networks of the same size.

Network parameters such as the shortest path length, clustering coefficient and closeness were calculated using the computer software Pajek, downloaded from: `http://vlado.fmf.uni-lj.si/pub/networks/pajek`. The shortest path length between two genes in a network reflects the overall network interconnectedness; the smaller the average shortest path length is, the higher chance that genes in this network could interact with the other genes. The clustering coefficient of a network is a measurement of the degree to which nodes in a network tend to cluster together; the larger the average clustering coefficient is, the more closely the genes are connected, forming modules. The closeness of a network measures the centrality of nodes within a network; nodes

that occur on shortest paths with other nodes have higher closeness than those that do not [153].

3.6 Acknowledgments

We firstly thank Tim Connallon for extremely helpful discussion related to epistasis. Discussions from Dr. Huifeng Jiang, Dr. Chris Myers, and Mr. Kaixiong Ye are also very much appreciated. This work was supported by the startup fund from Cornell University, NSF DEB-0949556 and NIH 1R01AI085286 awarded to Z.G

CHAPTER 4

A ROBUST AND EFFICIENT METHOD FOR ESTIMATING ENZYME COMPLEX ABUNDANCE AND METABOLIC FLUX FROM EXPRESSION DATA

A major theme in constraint-based modeling is unifying experimental data, such as biochemical information about the reactions that can occur in a system or the composition and localization of enzyme complexes, with high-throughput data including expression data, metabolomics, or DNA sequencing. The desired result is to increase predictive capability and improve our understanding of metabolism. The approach typically employed when only gene (or protein) intensities are available is the creation of tissue-specific models, which reduces the available reactions in an organism model, and does not provide an objective function for the estimation of fluxes. We develop a method, flux assignment with LAD (least absolute deviation) convex objectives and normalization (FALCON), that employs metabolic network reconstructions along with expression data to estimate fluxes. In order to use such a method, accurate measures of enzyme complex abundance are needed, so we first present an algorithm that addresses quantification of complex abundance. Our extensions to prior techniques include the capability to work with large models and significantly improved run-time performance even for smaller models, an improved analysis of enzyme complex formation, the ability to handle large enzyme complex rules that may incorporate multiple isoforms, and either maintained or significantly improved correlation with experimentally measured fluxes. FALCON has been implemented in MATLAB and ATS, and can be downloaded from: `https://github.com/bbarker/FALCON`. ATS is not required to compile the software, as intermediate C source code is available. FALCON requires use of the COBRA Toolbox, also implemented in MATLAB.

4.1 Introduction

FBA (flux balance analysis) is the oldest, simplest, and perhaps most widely used linear constraint-based metabolic modeling approach [2, 154]. FBA has become extremely popular, in part, due to its simplicity in calculating reasonably accurate microbial fluxes or growth rates (e.g. Schuetz et al. 7, Fong and Palsson 18); for many microbes, a simple synthetic environment where all chemical species are known suffices to allow proliferation, giving fairly complete constraints on model inputs. Additionally, it has been found that their biological objectives can be largely expressed as linear objectives of fluxes, such as maximization of biomass [7]. Neither of these assumptions necessarily hold for mammalian cells growing *in vitro* or *in vivo*, and in particular the environment is far more complex for mammalian cell cultures, which have to undergo gradual metabolic adaptation via titration to grow on synthetic media [155]. Recently, there have been many efforts to incorporate both absolute and differential expression data into metabolic models [156]. The minimization of metabolic adjustment (MoMA; Segrè et al. 157) algorithm is the simplest metabolic flux fitting algorithm, and it can be extended in order to allow the use of absolute expression data for the estimation of flux [1], which is the approach taken in this study.

The MoMA method is framed as a constrained least-squares optimization problem, is typically employed to calculate the flux vector of an *in silico* organism after a mutation by minimizing the distance between the wild-type flux and the mutant flux. The biological intuition is that the organism has not had time to adapt to the restricted metabolic capacity and will maintain a similar flux to the wild-type (WT) except where the perturbations due to the mutation dictate necessary alterations in fluxes [24]. Suppose $\mathbf{a}$ is the WT flux vector obtained by an optimization procedure such as FBA, empirical measurements, or a combination of these. For an undetermined flux vector $\mathbf{v}$ in a model

with N reactions the MoMA objective can be expressed as

$$\text{minimize} \sum_{i=1}^{N}(v_i - a_i)^2 \tag{4.1}$$

subject to the stoichiometric constraints $\mathbf{Sv} = \mathbf{0}$ where $\mathbf{v} = (v_1, \ldots, v_N)^T$ and $\mathbf{S}$ is the stoichiometric matrix (rows correspond to metabolites, columns to reactions, and entries to stoichiometric coefficients). Constant bounds on fluxes are often present, such as substrate uptake limits, or experimental $V_{\max}$ estimates, so we write these as the constraints $\mathbf{v}_{lb} \preceq \mathbf{v} \preceq \mathbf{v}_{ub}$. The objective may be equivalently expressed in the canonical quadratic programming (QP) vector form as min. $\frac{1}{2}\mathbf{v}^T\mathbf{v} - \mathbf{a}^T\mathbf{v}$. This assumes that each a_i is measured, but it is also possible and sometimes even more useful to employ this objective when only a subset of the a_i are measured (if a_i is not measured for some i, then we omit $(v_i - a_i)^2$ from the objective). In metabolomics, for instance, it is always the case in experiments with labeled isotope tracers that only a relatively small subset of all fluxes are able to be estimated with metabolic flux analysis (MFA; Shestov et al. 2). Combining MoMA with MFA provides a technique to potentially estimate other fluxes in the network.

A variant of MoMA exists that minimizes the absolute value of the difference between a_i and v_i for all known a_i. To our knowledge, the following linear program is the simplest version of linear MoMA, which assumes the existence of a constant flux vector **a**:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{N} d_i \\
\text{subject to} \quad & \mathbf{Sv} = \mathbf{0} \\
& \mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub} \\
\forall i: \quad & -d_i \leq v_i - a_i \leq d_i \\
& d_i \geq 0
\end{aligned}
\tag{4.2}
$$

The d_i are just the distances from *a priori* fluxes to their corresponding fitted fluxes. Linear MoMA has the advantage that it is not biased towards penalizing large magnitude fluxes or under-penalizing fluxes that are less than one [24, 158]. Additionally, linear programs are often amenable to more alterations that maintain convexity than a quadratic program [158].

We wish to apply MoMA to expression data rather than flux data, but there are two primary problems that must be tackled. First, we must quantify enzyme complex abundance as accurately as possible given the gene expression data. Although there is not a one-to-one correspondence between reactions and enzyme complexes, the correspondence is much closer than that between individual genes and metabolic reactions. In the first part of this work, we employ an algorithm that can account for enzyme complex formation and thus quantify enzyme complex abundance. Second, we must fit real-valued variables (fluxes) to non-negative data (expression), which is challenging to do efficiently. To accomplish this, we build on the original MoMA objective, which must be altered in several ways (also discussed in Lee et al. 1, which lays the groundwork for the current method). We develop automatic scaling of expression values so that they are comparable to flux units obtained in the optimization routine. This can be an advantage over the prior method as it no longer requires the manual choice of a flux and complex abundance pair with ratio that is assumed to be representative of every such pair in the system. Related to this, we also implement the sharing of enzyme com-

plex abundance between the reactions that the complex catalyzes, rather than assuming there is no competition between reactions catalyzed by the same complex. Reaction direction assignment enables comparison of fluxes and expression by changing fluxes to non-negative values. We show that batch assignment, rather than serial assignment [1] of reaction direction can greatly improve time efficiency. Finally, we employ several sensitivity analyses and performance benchmarks so that users of the FALCON method and related methods may have a better understanding of what to expect in practice.

4.2 Methods

Most genome-scale models have attached Boolean (*sans* negation) gene rules to aid in determining whether or not a gene deletion will completely disable a reaction. These are typically called GPR (gene-protein-reaction) rules and are a requirement for FALCON; their validity, like the stoichiometric matrix, is important for generating accurate predictions. Also important are the assumptions and limitations for the process of mapping expression data to complexes so that a scaled enzyme complex abundance (hereafter referred to as complex abundance) can be estimated. We address these in the next section and have attached a flow chart to illustrate the overall process of mapping expression of individual genes to enzyme complexes within the greater context of flux estimation (Figure 4.1). We employ an algorithm for this step—finding the minimum disjunction—for estimating complex abundance as efficiently and as accurately as possible given the assumptions (Section C.2).

Consideration of constraint availability, such as assumed reaction directions and nutrient availability, is crucial in this type of analysis. In order to work with two sets of constraints with significantly different sizes in yeast, we wrote the MATLAB func-

tion `removeEnzymeIrrevs` to find all enzymatic reactions in a model that are annotated as reversible but are constrained to operate in one direction only. The script then changes the bounds to allow flux in both directions. The function `useYN5irrevs` copies the irreversible annotations found in Yeast 5.21 [1] to a newer yeast model, but could in principle be used for any two models; by default, this script is coded to first call `removeEnzymeIrrevs` on both models before copying irreversible annotations. Application of these scripts removes 853 constraints in Yeast 5.21 and 1,723 constraints in Yeast 7. Despite the significant relaxation in constraints, since nutrient uptake constraints are unaffected, FBA only predicts a 1.28% increase in growth rate in the minimally constrained Yeast 7 model. However, in FALCON, we are no longer optimizing a sink reaction like biomass, and this relaxation in internal constraints proves to be more important. Constraint sets for Human Recon 2 are described in Figure C.4.

4.2.1 Estimating enzyme complex abundance

Given the diversity and availability of genome-scale expression datasets, either as microarray or more recently RNA-Seq, it could be useful to gauge the number of enzyme complexes present in a cell. A recent study found that only 11% of annotated *Drosophila* protein complexes have subunits that are co-expressed [159], so it cannot be assumed that any given protein subunit level represents the actual complex abundance. We formalize a model for enzyme complex formation based on GPR rules that are frequently available in genome-scale annotations.

The original expression to complex abundance mapping procedure [1] performed a direct evaluation of GPR rule expression values—replacing gene names with their expression values, ANDs with minimums, and ORs with sums, without altering the

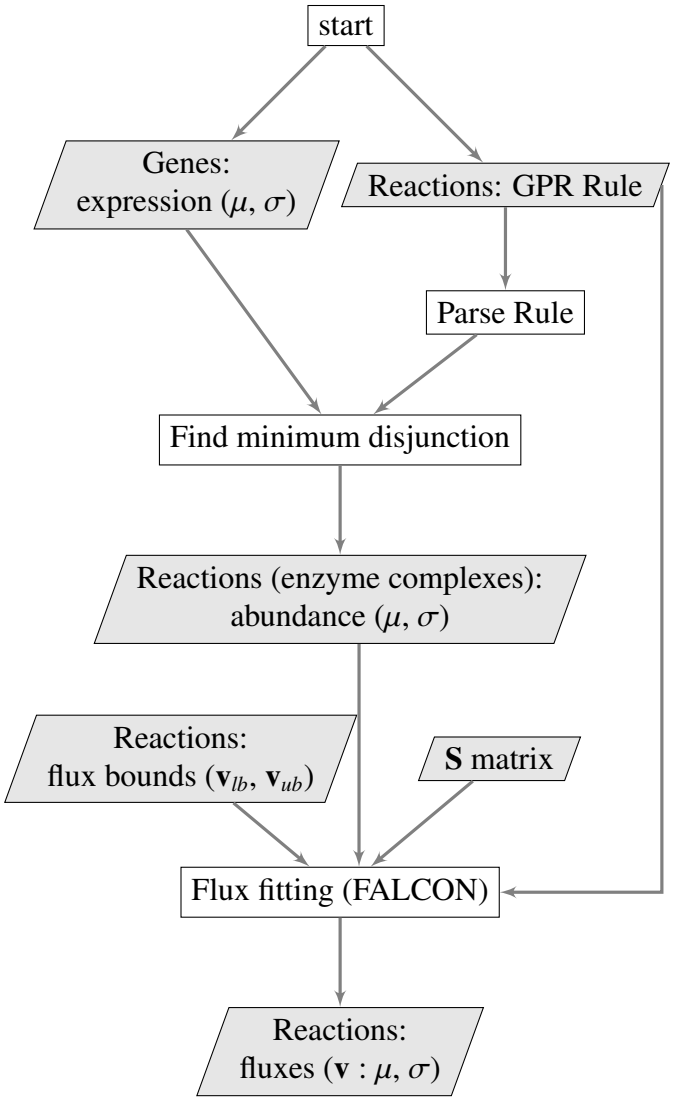


Figure 4.1: Flowchart illustrating the two algorithms used in this paper. The process of estimating enzyme complex abundance is displayed in detail, whereas the flux-fitting algorithm (FALCON) is illustrated as a single step for simplicity. First, for each gene in the model with available expression data, the mean and (if available) standard deviation or some other measure of uncertainty are read in. Gene rules (also called GPR rules) are also read in for each enzymatic reaction. The reaction rules are parsed and the minimum disjunction algorithm is applied, making use of the gene's mean expression. Next, the estimated and unitless enzyme complex abundance and variance are output for each enzymatic reaction. Finally, flux fitting with FALCON (Algorithm 1) can be applied, and requires the model's stoichiometry and flux bounds. The final output has the option of being a deterministically estimated flux, or a mean and standard deviation of fluxes if alternative optima are explored.

Boolean expression of the GPR rule in any way. Below we illustrate a problem that can occur with this mapping where some genes' expression levels may be counted more than once.

The r_i are different reaction rules and the e_i are the corresponding estimated complex abundance levels. Lower case letters are shorthand for the expression level of the corresponding gene ID in uppercase; for example, $a = \mathrm{E(A)}$, where $\mathrm{E(A)}$ is the expression of gene A.

$$\begin{aligned} r_1 &:= \text{[A and B] or [A and C]} \rightarrow e_1 = \min(a, b) + \min(a, c) \\ r_2 &:= \quad \text{[A and (B or C)]} \quad \rightarrow e_2 = \quad \min(a, b + c) \end{aligned} \tag{4.3}$$

Supposing A is the minimum, then if we just evaluate r_1 directly (a rule in disjunctive normal form, or DNF), A will be counted twice. Rules with sub-expressions in DNF are frequently encountered in practice, but directly evaluating them can lead to erroneous quantification.

Another possibility is partitioning expression among multiple occurrences of a gene in a rule. For instance, in r_1 above, we could evaluate it as $e_1 = \min(\frac{a}{2}, b) + \min(\frac{a}{2}, c)$ to account for the repeated use of a. However, other potential issues aside, we can see that this can cause problems rather quickly. For instance, suppose $b = a$ and $c = 0$; then $\min(a,b + c) = b = a$ appears to be correct, not $\min(\frac{a}{2}, b) + \min(\frac{a}{2}, c) = \frac{a}{2} + 0$. From this example, we can see that conversion to conjunctive normal form (CNF; Russell and Norvig 160), as in r_2 appears to be a promising prerequisite for evaluation.

4.2.2 The min-disjunction algorithm estimates enzyme complex abundance

In Section C.2, we show that converting a rule to CNF is a sound method to aid in the estimation of enzyme complex abundance. The minimum disjunction algorithm is essentially just the standard CNF conversion algorithm [160], with the implementation caveat that a gene that is in disjunction with itself should be reduced to a literal. We've found that this makes the CNF conversion algorithm tractable for all rules and prevents double counting of gene expression. Conversion to CNF and selection of the minimum disjunction also removes redundant genes from the complex (e.g. holoenzymes; see Assumption 8). Biologically, selecting the minimum disjunction effectively finds the *rate-limiting* component of enzyme-complex formation. After conversion to CNF, the minimum disjunction algorithm substitutes gene-expression values as described in Lee et al. 1 and evaluates the resulting arithmetic expression. Another new feature of our approach is the handling of missing gene data. If expression is not measured for a gene in a GPR rule, the rule is modified so that the missing gene is no longer part of the Boolean expression. For instance, if data is not measured for gene B in [A and (B or C)] then the rule would become [A and C]. This prevents penalization of the rule's expression value in the case that the missing gene was part of a conjunction, and it also assumes there was no additional expression from the missing gene if it is in a disjunction.

Although conversion to CNF may be intractable for some expressions [160], we tested our implementation of the algorithm on three of the most well-curated models which contain some of the most complex GPR rules available. These models are for *E. coli* [161], yeast [162], and human [63]. In all cases, the rules were converted to CNF in less than half a second, which is far less than the typical flux fitting running time from Algorithm 1.

Using the minimum disjunction method results in several differences from direct substitution and evaluation in yeast GPR rules. When data completely covers the genes in the model (e.g. Lee et al. 1), complex abundance tends to have few differences in yeast regardless of the evaluation method (25 rules; 1.08% of all rules for Yeast 7). This number goes up significantly in Human Recon 2 [63] due to more complex GPR rues (935 rules; 22% of all rules). For the human model, we could not find any data set that covered every gene, so instead random expression data roughly matching a power law was used to generate this statistic. If we use proteomics data for yeast and human models, the algorithmic variation in how missing gene data is handled causes some additional increase in differences [163, 164]. For proteomics, in the Yeast 7 model 205 rules (8.87% of all rules) differed, and in Human Recon 2, 1002 rules (23.57% of all rules) differed. We can see that for yeast, the changes in flux attributed to enzyme abundance evaluation can be relatively small for data with 100% gene coverage, but can be significant in Human (Figure C.1).

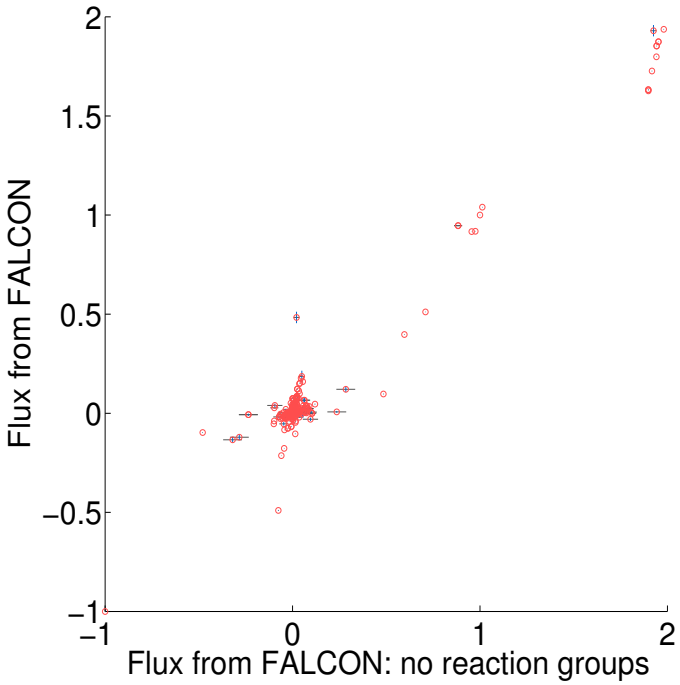
4.3 The FALCON algorithm

Prior work that served as an inspiration for this method used Flux Variability Analysis (FVA) to determine reaction direction [1]. Briefly, this involves two FBA simulations per reaction catalyzed by an enzyme, and as the algorithm is iterative, this global procedure may be run several times before converging to a flux vector. We removed FVA to mitigate some of the cost, and instead assign flux direction in batch; while it is possible that the objective value may decrease using this approach, this is not an issue since the objective function increases to include more irreversible fluxes at each iteration, and the objective value of a function with more fluxes should supersede the importance of one with fewer fluxes.

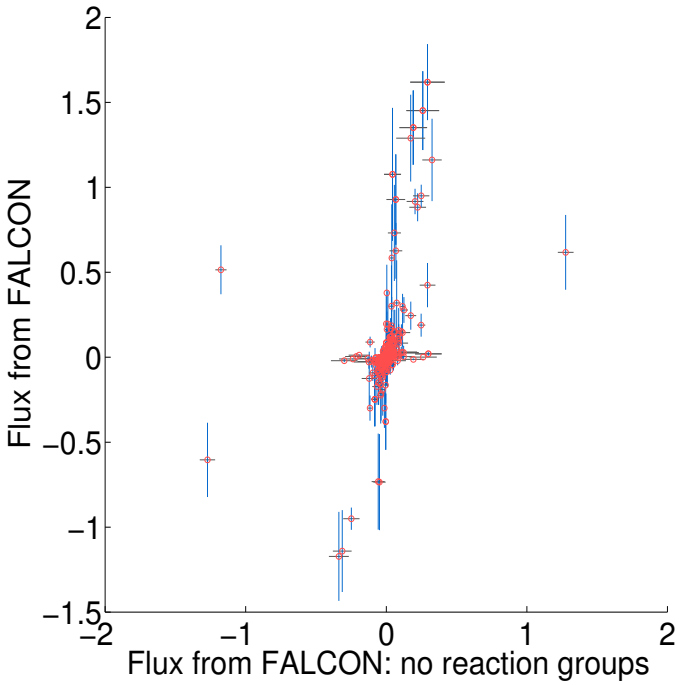
One major advance in our method is the consideration of enzyme complexes sharing multiple reactions, which we call reaction groups. This is done by partitioning an enzyme complex's abundance across its reactions by including all reactions associated to the complex in the same constraint. Both minimally and highly constrained models (Section 4.4.2) show some fluxes with significant differences depending on the use of group information, particularly in the minimally constrained model (Figure 4.2). We now discuss the algorithm in detail, including several other important features, including automatic scaling of expression.

To make working with irreversible fluxes simpler, we convert the model to an irreversible model, where each reversible flux v_j in the original model is split into a forward and a backward reaction that take strictly positive values: $v_{j,f}$ and $v_{j,b}$. We also account for enzyme complexes catalyzing multiple reactions by including all reactions with identical GPR rules in the same residual constraint; indexed sets of reactions are denoted R_i and their corresponding estimated enzyme abundance is e_i. Figure 4.2 shows the difference in Algorithm 1 when we do not use reaction group information. The standard deviation of enzyme abundance, σ_i, is an optional weighting of uncertainty in biological or technical replicates.

We employ a normalization variable n in the problem's objective and flux-fitting constraints to find the most agreeable scaling of expression data. The linear fractional program shown below can be converted to a linear program by the Charnes-Cooper transformation [158]. To avoid the need for fixing any specific flux, which may introduce bias, we introduce the bound $\sum_{j \mid \exists i \text{ s.t. } j \in R_i} \left|v_j\right| \geq V_{lb}^{\Sigma}$. This guarantees that the optimization problem will yield a non-zero flux vector. As an example of how this can be beneficial, this means we do not need to measure any fluxes or assume a flux is fixed to achieve good results; though this does not downplay the value of obtaining



(a)



(b)

Figure 4.2: Comparison of setting FALCON to use no reaction group information (x-axis) versus with group information (y-axis; default FALCON setting) for both the highly constrained Yeast 7 model **(a)** and the minimally constrained Yeast 7 model **(b)**. Error bars with length equal to one standard deviation are shown for both approaches as a result of alternative solutions in FALCON.

experimentally-based constraints on flux when available (Figure C.2).

The actual value of V_{lb}^{Σ} is not very important due to the scaling introduced by n, and we include a conservatively small value that should work with any reasonable model. However, for numeric reasons, it may be best if a user chooses to specify a value appropriate for the model. Similarly, if any fluxes are known or assumed to be non-zero, this constraint becomes unnecessary. To keep track of how many reactions are irreversible in the current and prior iteration, we use the variables $rxns_{irrev}$ and $rxns_{irrev,prior}$. The algorithm terminates when no reactions are constrained to be exclusively forward or backward after an iteration.

Algorithm 1 and the method in Lee et al. 1 are both non-deterministic. In the first case, Algorithm 1 solves an LP during each iteration, and subsequent iterations depend on the LP solution, so that alternative optima may affect the outcome. In the latter case, alternative optima of individual LPs is not an issue, but the order in which reactions assigned to be irreversible can lead to alternative solutions. However, we found that the variation due to this stochasticity is typically relatively minor, particularly in cases where the model is more heavily constrained (Figs. C.1 and C.2).

4.4 Results and Discussion

4.4.1 Performance benchmarks

Using the same yeast exometabolic and expression data employed for benchmarking in the antecedent study [1] that included an updated version of the Yeast 5 model [165] and the latest yeast model [162], we find that our algorithm has significant improvements

Algorithm 1 FALCON

INPUT: $\{R_i : i$ an index for a unique enzyme complex where
$R_i = \{j :$ complex i catalyzes reaction $j\}\}$
INPUT: enzyme abundances (mean: e_i, standard deviation: σ_i)
INPUT: model (**S** matrix, $\mathbf{v}_{lb}$, $\mathbf{v}_{ub}$)
$u_{\min} \leftarrow \min_j \{V_{j,\max} : V_{j,\max} > 0\}$ where $V_{j,\max} = \max\left(\left|v_{lb,j}\right|, \left|v_{ub,j}\right|\right)$.
$V_{lb}^{\Sigma} \leftarrow u_{\min} |\{j : \exists i \text{ s.t. } j \in R_i\}|$
$rxns_{irrev} \leftarrow$ number of reactions (j) such that either $v_{ub,j} > 0$ or $v_{lb,j} < 0$, but not both.
for all i **do**
Scale data to be of similar size for numeric stability:
$e_i \leftarrow \frac{e_i V_{lb}^{\Sigma}}{\sum_j e_i}$
$\sigma_i \leftarrow \frac{\sigma_i V_{lb}^{\Sigma}}{\sum_j e_i}$
end for
while $rxns_{irrev} > rxns_{irrev,prior}$ **do**
$rxns_{irrev,prior} \leftarrow rxns_{irrev}$
Call LP Solver (updates **v**):
minimize $\sum_i \frac{d_i}{n\sigma_i}$
s.t.
$\sum_{j \mid \exists i \text{ s.t. } j \in R_i} \left|v_j\right| \geq V_{lb}^{\Sigma}$
$\forall i : -d_i \leq \sum_{j \in R_i} (v_{j,f} + v_{j,b}) - ne_i \leq d_i$ where $v_j = v_{j,f} - v_{j,b}$
$d_i, v_{j,f}, v_{j,b} \geq 0$
$n > 0$
for all $\left\{j \mid v_{j,f} + v_{j,b} > 0, v_{j,f} \neq v_{j,b}\right\}$ **do**
Constrain the smaller of $v_{j,f}$ and $v_{j,b}$ to be 0.
$rxns_{irrev} \leftarrow rxns_{irrev} + 1$
end for
end while
OUTPUT: **v**

Table 4.1: Performance of FALCON and other CBM methods for predicting yeast exometabolic fluxes in two growth conditions with highly (HC) and minimally (MC) constrained models **(a)** and associated timing analysis **(b)**. For Lee et al. and FALCON methods, the mean time for a single run of the method is listed; all other methods did not have any stochasticity employed. Values are shown in two significant figures. Method descriptions can be found in Lee et al. 1.

(a)	Max. μ	Model	Experimental	Standard FBA	Fitted FBA	GIMME	iMAT	Lee et al.	FALCON
	75 %	Yeast 5 MC	1	0.66	0.66	NaN	0.57	0.64	1
	75 %	Yeast 7 MC	1	0.66	0.66	0.68	0.66	0.66	0.98
	75 %	Yeast 5 HC	1	0.73	0.78	0.75	0.66	0.98	0.99
Pearson's r	75 %	Yeast 7 HC	1	0.70	0.70	0.80	0.66	0.98	0.99
	85 %	Yeast 7 MC	1	0.62	0.62	0.65	0.62	0.62	0.97
	85 %	Yeast 5 HC	1	0.88	0.89	0.9	0.81	0.99	0.99
	85 %	Yeast 7 HC	1	0.67	0.67	0.87	0.62	0.98	0.98
(b)	Max. μ	Model	Experimental	Standard FBA	Fitted FBA	GIMME	iMAT	Lee et al.	FALCON
	75 %	Yeast 5 MC	0	0.9	470	0.81	50	110	1.8
	75 %	Yeast 7 MC	0	1.9	3,100	2.1	12,000	600	5.6
	75 %	Yeast 5 HC	0	0.12	110	0.18	1.4	15	0.27
Time (s)	75 %	Yeast 7 HC	0	0.72	940	1.7	240	670	5.5
	85 %	Yeast 7 MC	0	2.3	3,100	3.8	14,000	610	4.6
	85 %	Yeast 5 HC	0	0.12	110	0.18	2.5	15	0.22
	85 %	Yeast 7 HC	0	0.70	110	2.5	100	530	5.9

in time efficiency while maintaining correlation with experimental fluxes, and is much faster than any similarly performing method (Table 4.1; Figure C.2). Timing for the human model also improved in FALCON; in a model with medium constraints and exometabolic directionality constraints, FALCON completed on average in 3.6 m and the method from Lee et al. 1 in 1.04 h. Furthermore, when we remove many bounds constraining the direction of enzymatic reactions that aren't explicitly annotated as being irreversible in prior work [1], we find that our formulation of the approach seems to be more robust than other methods.

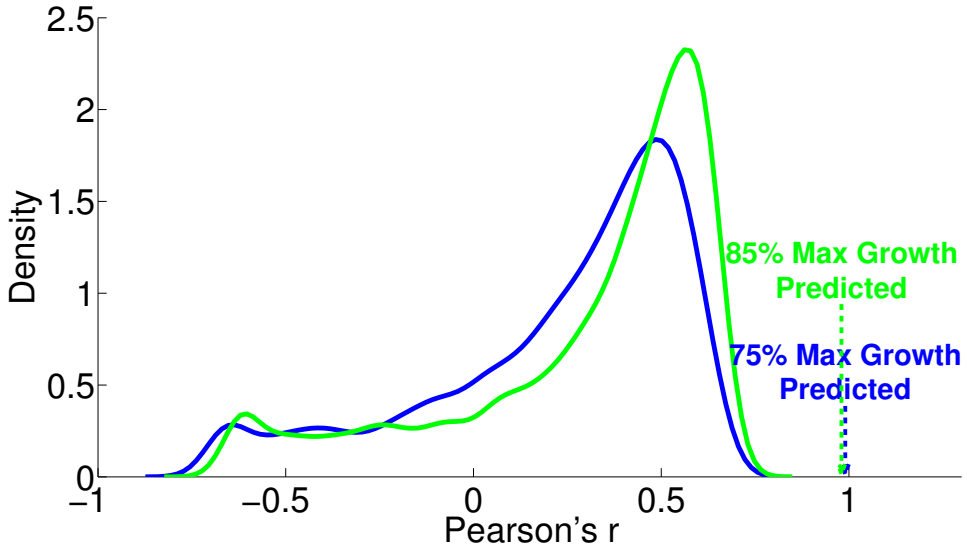
We see that the predictive ability of the algorithm does not appear to be an artifact; when FALCON is run on permuted expression data, it doesn't do as well as the actual expression vector (Figure 4.3). The full-sized flux vectors estimated from permuted expression as a whole also does not correlate well with the flux vector estimated from the actual expression data, but we notice that the difference is visibly larger in the minimally constrained model compared to the highly constrained model (Figure C.3). Rigidity

in the highly constrained model appears to keep most permutations from achieving an extremely low Pearson correlation, likely due to forcing fluxes through the same major pathways, but a rank-based correlation still shows strong differences.

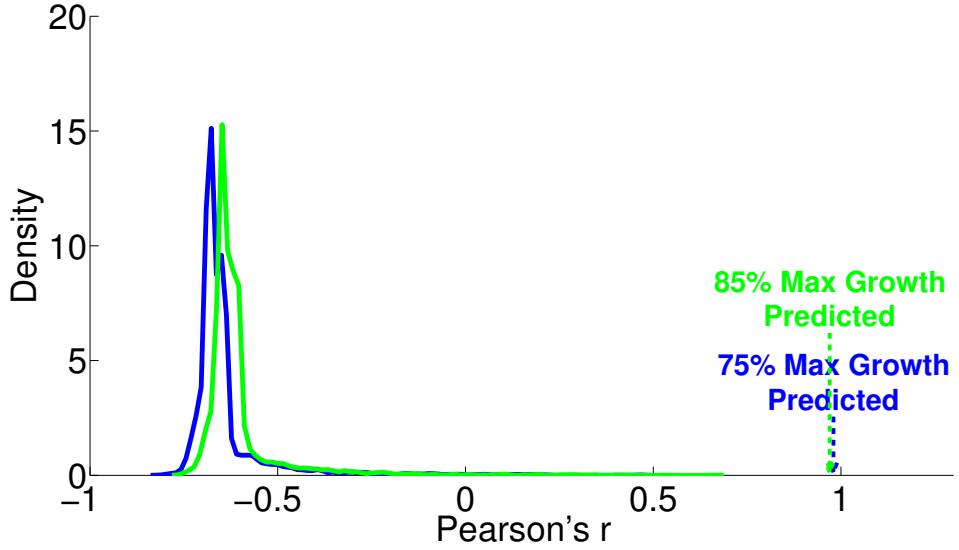
4.4.2 Sensitivity to expression noise

To understand the sensitivity of flux to expression, we multiply noise from multivariate log-normal distributions with the expression vector and see the effect on the estimated fluxes. For instance, correlation between two types of proteomics data yields a Pearson's $r = 0.7$ [164], corresponding to an expected $\sigma \approx 1.4$ and expected $r \approx 0.4$ for flux in our most highly constrained human model (Figure C.4). We find that enzymatic reaction directionality constraints influence the sensitivity of the model to expression perturbation (Figure 4.4.2). It is important to note that mere presence of the constraints does not help us determine the correct experimental fluxes when other classes of methods (e.g. FBA; Table 4.1) are used. Additionally, it is possible to obtain good predictions even without a heavily constrained model (Table 4.1).

With Human Recon 2, additional constraint sets supply some benefit, but even the most extreme constraint set does not compare to what is available in Yeast 7, which is also inherently constrained by the fact that yeast models will be smaller than comparable human models (Figure C.4). For mammalian models, more sophisticated means of constraint, such as enzyme crowding constraints [60], or using FALCON in conjunction with tissue specific modeling tools, may prove highly beneficial.



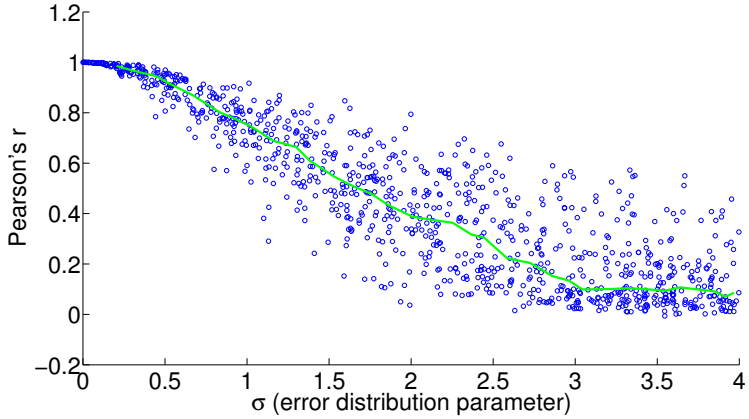
(a) highly constrained



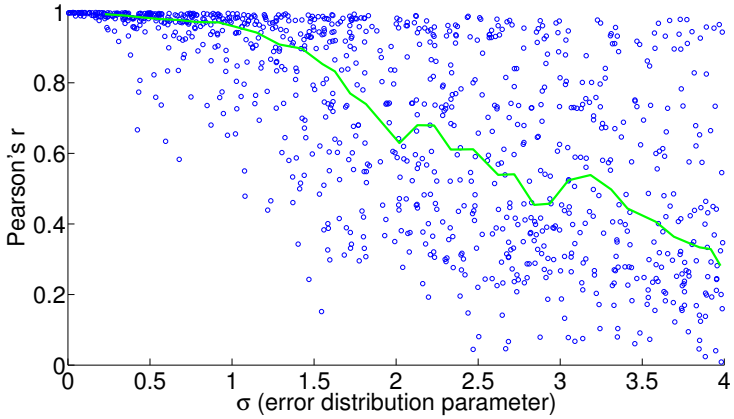
(b) minimally constrained

Figure 4.3: Kernel-smoothed PDFs of correlation between experimental fluxes and fluxes estimated from FALCON when all gene expression data points are permuted. Arrows mark the correlation when FALCON is run on the unpermuted expression data. Random correlations tend to be much more positive in the highly constrained model **(a)** than in the minimally constrained model **(b)**. 5,000 permutation replicates were performed in all cases.

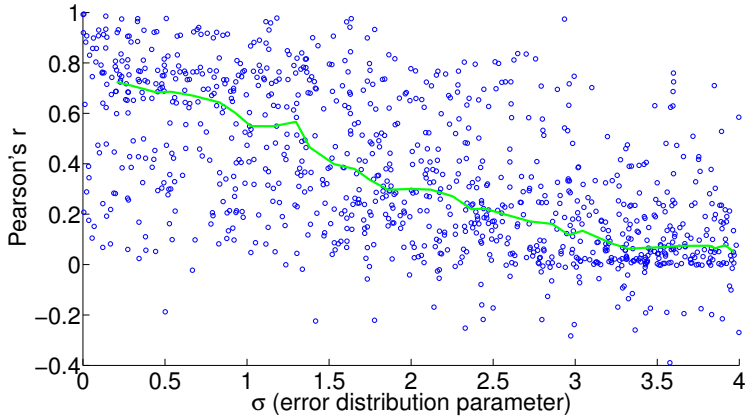
Figure 4.4: Correlation of perturbed enzyme abundance vectors and flux vectors with the associated unperturbed vector for the Yeast 7 model. The interval median correlation is shown in green. Noise sampled from a multivariate log-normal distribution with parameters $\mu = 1$ and σ (x-axis) is multiplicatively applied to the enzyme abundance vector, and the y-axis shows the Pearson correlation between the two vectors **(a)**. Similar plots show correlation between flux vectors estimated with FALCON using the same perturbed and unperturbed expression vectors **(b-c)**.



(a) enzyme complex abundance



(b) flux: highly constrained



(c) flux: minimally constrained

4.4.3 Flux estimates provides information beyond enzyme complex abundance

It is not an unreasonable hypothesis that fluxes would correlate well with their associated complex abundances. Indeed, the general principle needed for fitting fluxes to enzyme complex abundances is to assume the values would be correlated in the absence of other constraints (e.g. branch points that arise from the stoichiometry). More specifically, it should be the case that flux is proportional to enzyme complex abundance given ample availability of substrate, and that this proportionality constant does not vary too much between reactions. There are undoubtedly many exceptions to this rule, but it seems as though there may be some underlying evolutionary principles for it to work in this parsimonious fashion, as has been partly verified [166].

Aside from the obvious benefits of constraint-based methods also estimating fluxes for non-enzymatic reactions, and assigning a direction for reversible enzymatic reactions, we see that in general, our method does not predict a strong correlation between complex abundance and flux (Figure C.5). Recently it has been shown that many fluxes are not under direct control of their associated enzyme expression level [167], which gives experimental support to the idea that a network-based approach, such as that presented in this paper, may be useful in understanding how fluxes may be constrained by expression data. Chubukov et al. 167 also note that enzymes may be overexpressed in some cases, either for robustness or because of noise in transcriptional regulation. This will not usually be a problem in FALCON, unless entire pathways are overexpressed, which would be unusual as it would represent a seemingly large energetic inefficiency.

The present work doesn't attempt to use empirically obtained kinetic parameters to estimate $V_{\max}$, but this approach does not seem as promising in light of experimental

evidence that many reactions in central carbon metabolism tend to operate well below $V_{\max}$ [166]. Still, a better understanding of these phenomena may make it possible to improve flux estimation methods such as the one presented here, or more traditional forms of MFA [2] by incorporating enzyme complexation and kinetic information.

4.4.4 Increasing roles for GPR rules and complex abundance estimates

Still, complex abundance may have uses aside from being a first step in FALCON. The method presented here for complex abundance estimation can be used as a stand-alone method, as long as GPR rules from a metabolic reconstruction are present. For instance, it may not always be desirable to directly compute a flux. As an example, the relative abundance of enzyme complexes present in secretions from various biological tissues, such as milk or pancreatic secretions, may still be of interest even without any intracellular flux data. Perhaps more importantly, this approach to estimating relative complex levels can be employed with regulatory models such as PROM [31] or other regulatory network models that can estimate individual gene expression levels at time $t + 1$ given the state of the model at a time t.

GPR rules and stoichiometry may be inaccurate or incomplete in any given model. In fact, for the foreseeable future, this is a given. By using the GPR and not just the stoichiometry to estimate flux, it is possible that future work could make use of this framework to debug not just stoichiometry as some methods currently do (e.g. Reed et al. 168) , but also GPR rules. Hope for improved GPR rule annotation may come from many different avenues of current research. For instance, algorithms exist for reconstructing biological process information from large-scale datasets, and could be

tuned to aid in the annotation of GPR rules [169]. Flexible metabolic reconstruction pipelines such as GLOBUS may also be extended to incorporate GPR rules into their output, and in so doing, extend this type of modeling to many non-model organisms [170]. Another limitation that relates to lack of biological information is that we always assume a one-to-one copy number for each gene in a complex. Once more information on enzyme complex structure and reaction mechanism becomes available, an extension to the current method could make use of this information. Even at the current level of structure, we think it is evident that GPR rules should undergo some form of standardization; Boolean rules without negation may not always capture the author's intent for more complex purposes like flux fitting.

4.5 Conclusion

We have formalized and improved an existing method for estimating flux from expression data, as well as listing detailed assumptions in Table C.1 that may prove useful in future work. Although we show that expression does not correlate well with flux, we are still essentially trying to fit fluxes to expression levels. The number of constraints present in metabolic models (even the minimally constrained models) prevents a good correlation between the two. However, as with all constraint-based models, constraints are only part of the problem in any largely underdetermined system. We show that gene expression can prove to be a valuable basis for forming an objective, as opposed to methods that only use expression to further constrain the model by creating tissue-specific or condition-specific models [29, 30].

For better curated models, the approach described immediately finds use for understanding metabolism, as well as being a scaffold to find problems for existing GPR rules,

and more broadly the GPR formalism itself. The present results and avenues for future improvement show that there is much promise for using expression to estimate fluxes, and that it can already be a useful tool for performing flux estimation and analysis.

4.6 Acknowledgments

We thank Michael Stillman for discussion and reading the manuscript, and we are also grateful to Neil Swainston for discussions on various constraint-based modeling topics.

CHAPTER 5

EPISTATIC LANDSCAPES ARISING FROM ADAPTIVE MUTATIONS

Existing literature has only dealt with the simulation of strictly deleterious mutants rather than beneficial mutants in the constraint-based modeling literature, which is chiefly due to existing studies optimizing the fitness function, which leaves no room for improvement. In this study, we develop a constraint-based approach that can simulate beneficial, neutral, and deleterious mutations. We show that this simulation technique can be useful for understanding adaptive trajectories, and we develop a software library for the analysis of evolutionary paths. Our mechanistic model can reproduce the distribution of epistases between beneficial mutations that was observed in a data set and a population genetic model fit to the same data set, showing that our model behaves appropriately in this conext and may be a useful tool for further evolutionary analyses. Finally, in experimental data sets and in our simulations, slightly beneficial mutations are much more likely to have positive (synergistic) epistasis with other beneficial mutations, making their likelihood of becoming fixed higher than would be expected without considering epistatic effects.

5.1 Introduction

5.1.1 Adaptive mutations

Biologists have long wondered the extent to which evolution occurs due to nearly neutral and slightly deleterious mutations, or adaptive mutations, or more complex situations involving these types of mutations as well as their epistatic interactions [171, 172].

The occurrence of beneficial mutations and how they affect adaptation is currently an area of active interest in evolutionary biology [173, 174]. Although much focus in the past has been placed on deleterious mutations because of their prevalence in nature and disease, it is ultimately beneficial mutations that are responsible for adaptive evolution.

Since beneficial mutations are relatively rare, and since combining multiple mutants in the lab is increasingly difficult for the more mutations that are to be combined, an *in silico* analysis can shed light on what may be expected in adaptive evolution.

5.1.2 The need for a new modeling framework

Clearly, using FBA with the growth objective alone is not enough–we only ever get the optimum for our fitness objective. A way to circumvent this issue is to associate a feature of the system (e.g. flux into biomass) as the fitness while optimizing some other objective. This latter mechanism is not generally used as most models tend to be rather under-constrained as is. However, as we've seen, the FALCON method (Chapter 4) and other fitting methods like MoMA provide a way to use high-throughput data to introduce many additional constraints to the system.

Having a systems tool that can work with models of particular organisms will not only add another tool in the computational evolution and population genetics arsenal, but also in applied fields such as evolutionary engineering of microbial engineering, and understanding which gene mutant combinations which may be most advantageous for a cancer cell population.

5.2 Results

5.2.1 Weighted MoMA-FBA objectives

Minimization of metabolic adjustment (MoMA), along with Flux Balance Analysis (FBA), has proven successful in simulating growth rates and predicting in silico fluxes. Here we discuss a weighted objective approach that combines both objectives.

The typical MoMA problem is framed as a least-squares optimization problem and is typically employed to calculate the flux vector of an *in silico* organism after a mutation [23]. The biological intuition is that the organism has not had time to adapt to the restricted metabolic capacity and will maintain a similar flux to the wild-type (WT). If $\mathbf{a}$ is the WT flux vector obtained by an optimization procedure, such as min-norm FBA (flux balance analysis), then in a model with N reactions the optimization objective can be expressed as

$$\text{minimize } \sum_{i=1}^{N}(v_i - a_i)^2$$

subject to the stoichiometric constraints $\mathbf{Sv} = \mathbf{0}$ where $\mathbf{S}$ is the stoichiometric matrix and $\mathbf{v} = (v_1, \ldots, v_N)^T$ is the undetermined flux vector. Additional constant bounds on fluxes are often present, such as substrate uptake limits, so we write $\mathbf{v}_{lb} \preceq \mathbf{v} \preceq \mathbf{v}_{ub}$. The objective may be equivalently expressed in the canonical quadratic programming (QP) form

$$\text{minimize } \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x}$$

as

$$\text{minimize } \frac{1}{2}\mathbf{v}^T\mathbf{v} - \mathbf{a}^T\mathbf{v}.$$

When we try the standard MoMA procedure in a new environment, several things become clear. For instance, switching the energy and carbon source from glucose to

ethanol will result in zero growth because the biomass is relatively small compared to other fluxes numerically. This is easily fixed by adding a weight to the biomass, though there are also certain numerical issues related to this with at least some solvers (e.g. MOSEK) that require an inverse weight on all other fluxes.

To accommodate these weights, let w be a positive scalar weight, and let $f(w)$ and $g(w)$ be two scalar weight functions such that $f(w) > 0$ and $g(w) > 0$ for all $w \geq 1$. We also assume $f(w)$ is monotonically increasing and $g(w)$ is monotonically decreasing. While $g(w)$ is employed to achieve the inverse weighting mentioned above, $f(w)$ is used to achieve gradual regularization as w increases. Biologically, regularization is important as it allows efficiency in enzyme synthesis to be modeled, which is typically done in FBA by requiring a min-norm flux. We incorporate these weights into the objective:

$$\text{minimize} \sum_{i=1,i\neq b}^{N} g(w)(v_i - a_i)^2 - wv_b + \sum_{i=1,i\neq b}^{N} f(w)v_i^2$$

Index b corresponds to the growth or biomass pseudo-reaction. In the simulations below, $g(w) = \frac{1}{\sqrt{w}}$. For some simulations we haven't yet used a nonzero $f(w)$, but an example we have used is $f(w) = \frac{\log(\log w+1)}{C}$ where C is a scaling constant.

Express this objective as

$$\text{minimize } \frac{1}{2}\mathbf{v}^T\mathbf{Q}\mathbf{v} - \tilde{\mathbf{a}}^T\mathbf{v}$$

where $b = N$ for convenience, $\tilde{\mathbf{a}} = g(w)\mathbf{a}$ with the Nth entry replaced by w, and

$$\mathbf{Q} = \begin{pmatrix} f(w) + g(w) & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & f(w) + g(w) & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}.$$

For convenience, informally denote this objective function as $\mathcal{M}(w, d_1, \ldots, d_m)$ where the optional d_i represent possible mutations to the model.

The difference in smoothness between weighted quadratic MoMA and linear MoMA can be illustrated by considering the growth rate to be a function of the weight placed on the biomass objective (Figure 5.1). An advantage of quadratic MoMA is that it simulates a continuous range of fitnesses potentially matching any required fitness level exactly. Quadratic programming does not admit alternative optima either, if the objective is convex, as in MoMA. However, these benefits are not necessarily always biologically relevant; discrete mutations may give a discrete jump in fitness. There is no obvious reason why having a unique optima would be biologically preferred; in fact, when compared to simple objectives like MoMA or FBA, most biological systems appear to operate suboptimally [7]. Weighted linear MoMA can model these discrete transitions in flux state, and often this is reflected in discrete jumps in fitness, as seen above.

In the above figure, the difference in fitness between weighted quadratic MoMA and linear MoMA can be seen. An advantage of quadratic MoMA is that it simulates a continuous range of fitnesses, potentially matching any required fitness level exactly. However, this is not necessarily always biologically relevant; in reality, discrete mutations may give a discrete jump in fitness. Weighted linear MoMA models these discrete transitions in flux state, and often this is reflected in discrete jumps in fitness, as seen above.

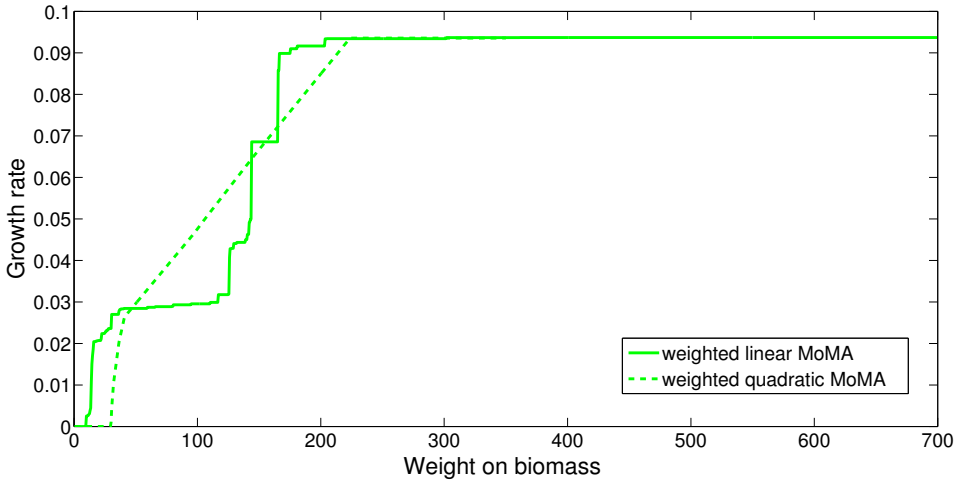
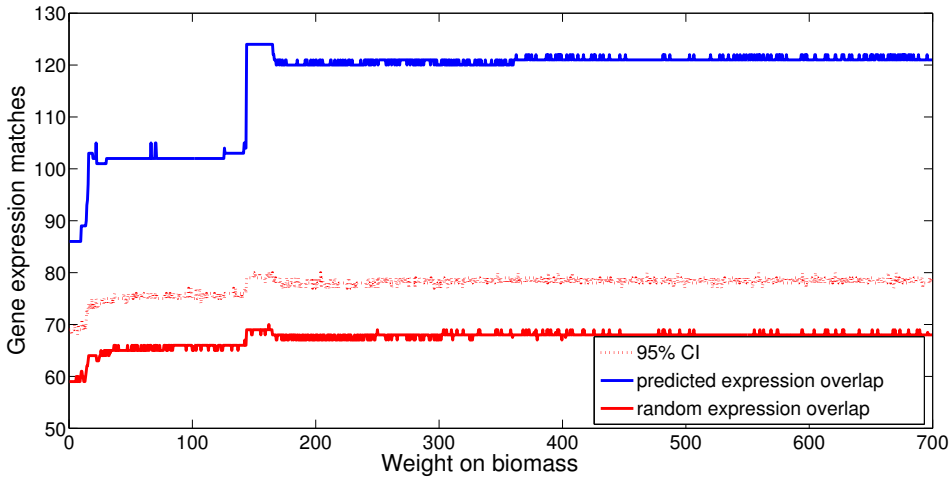


Figure 5.1: Weight on the biomass component of an objective (x-axis) influences the growth rate, where the other objective component is a MoMA objective that tries to minimize the flux difference with an ancestral environment.

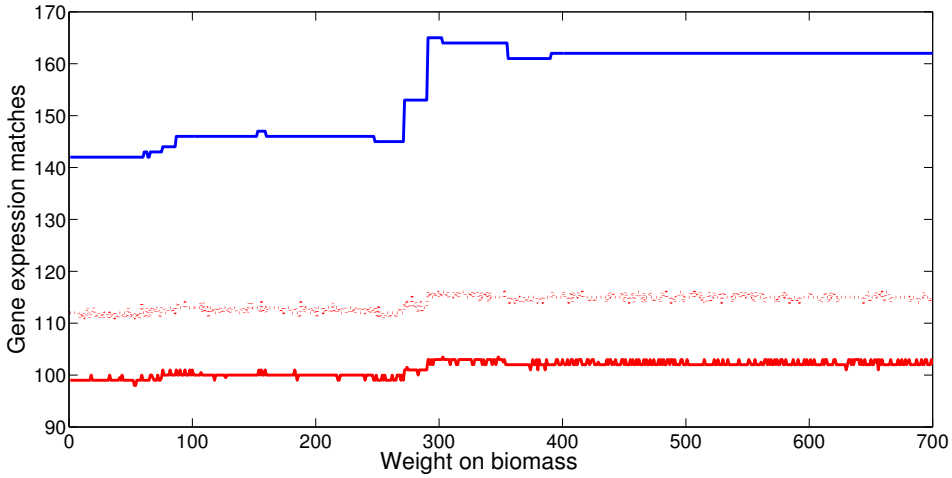
Weighted MoMA predicts expression states

Using tiling array data for s288c in YPD and YPE conditions from Xu et al. [175], we compare the number of genes expected to have a two-fold or greater change in expression from YPD to YPE to the fluxes mapped to genes also having greater than two-fold change in the same environmental transition. The x-axis represents the weight on growth. The number of genes having more than two-fold flux and expression change in the model and and in the experiment are shown in blue. The red and red dotted lines represent the random expectation and 95% CI for expression agreement for a random expression vector with the genes belonging to the model and experimental dataset. The random expression vector has the same number of two-fold up-regulated and two-fold down-regulated genes as the real relative expression vector.

Interestingly, the weight and corresponding fitness that agrees most with the experimental predictions is sub-optimal in both the quadratic and linear cases, suggesting that the s288c strain was not fully adapted to the YPE environment when the expression was



(a) Linear MoMA



(b) Quadratic MoMA

Figure 5.2: Number of yeast genes that are at least two-fold up-regulated or down-regulated when going from YPD to YPE media with matching predictions from weighted MoMA (blue line) and the associated 95% confidence interval (red dotted line) around the prediction to random chance (red line).

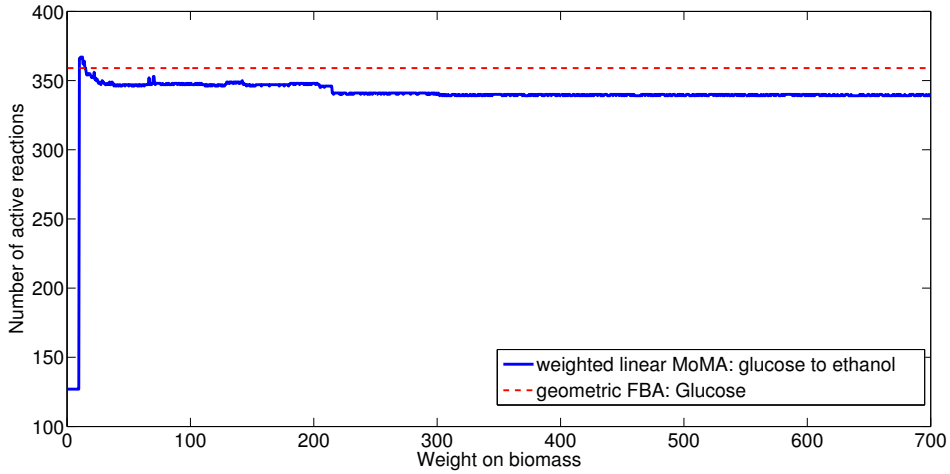


Figure 5.3: The number of active reactions (reactions with non-zero flux) plotted against the weight on the biomass component of a weighted linear MoMA objective (blue). An flux that maximizes biomass production that is centered among alternative optima is shown for comparison (red dotted line).

assayed. To compare directly with geometric FBA (a minimal $\boldsymbol{L}^1$-norm solution), we found that geometric FBA had 123 concurring genes whereas weighted linear MoMA had 124; similarly, a minimal $\boldsymbol{L}^2$-norm FBA solution in Ethanol had 162 agreeing solutions and weighted quadratic MoMA had 165, showing that weighted MoMA can do at least as good as the gold standards for *de novo* FBA solutions in predicting accurate fluxes.

Simulated adaptation induces complexity in sub-optimal solutions

Previous work explored the complexity of optimal and sub-optimal flux vectors, finding that generally the sub-optimal solutions have more active fluxes than the optimal solutions [176]. We found a similar trend with weighted MoMA where, once growth is non-zero, the initial flux state has more active reactions than in the ancestral environment (glucose; red dashed line) or than in the more adapted stages in the current (ethanol) environment (Figure 5.3).

5.2.2 Adaptive mutations with objective weights

Aside from the problem of separating the fitness function from the optimization objective function, there is the issue of combining traditional flux restriction mutations, which are known as *hard constraints*, which may result in an unsolvable system — an almost certainly undesirable effect of this mutation modeling formalism. Instead, it would be better if mutations could be modeled as *soft constraints*. Concretely, whereas hard constraints are enacted in the actual constraints of the optimization problem, soft constraints merely change the objective. This means that multiple soft constraints combined together under some mutational model would be compatible in the sense that they wouldn't unexpectedly result in an unsolvable system. When performing growth optimization in FBA, normally only the biomass pseudo-reaction has a non-zero (positive) entry. Nonzero values for any other entry could only decrease the flux. In weighted MoMA or FALCON, nonzero coefficients for any enzymatic reaction could prove potentially beneficial, as they may push the system in a direction that is more in line with growth optimization and less in line with MoMA or the FALCON flux-fitting objective.

FALCON and weighted MoMA provide two possible avenues for soft constraints: expression level and expression variation. However, it is not clear yet what expression variation really means, so further investigation is necessary. Furthermore, using a hard constraint model for mutants, as opposed to a soft constraint like weights on fluxes, appears to reduce the amount of non-trivial interactions: in the hard-constraint method one highly beneficial mutation may exactly subsume the mutation of a less beneficial mutant by forcing flux through the reaction; this happens to a certain extent in the weighted model as well, but not nearly as frequently.

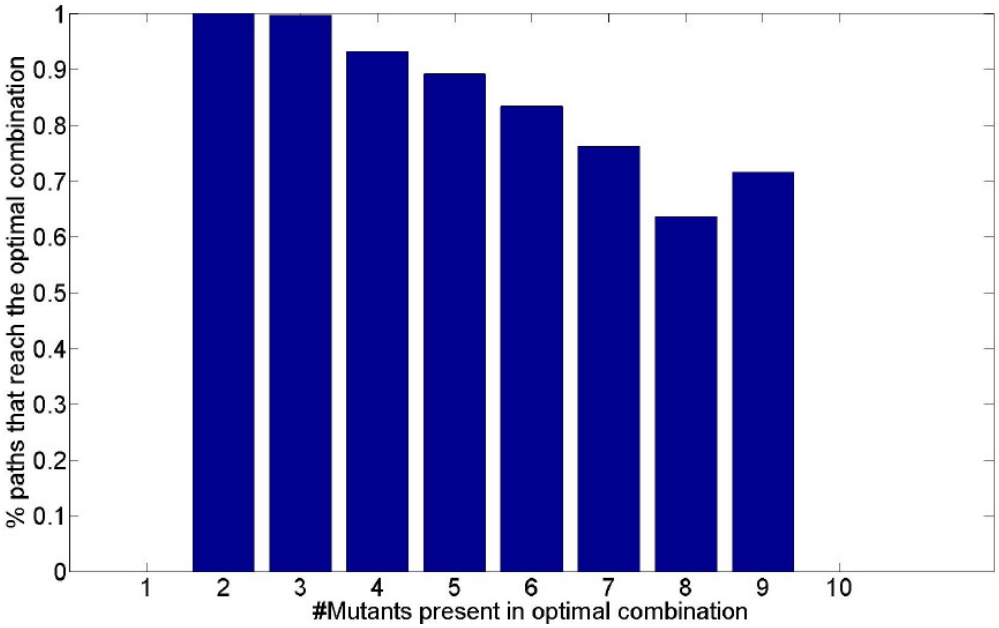
An example for both the quadratic and linear cases exemplifies the difference in smoothness obtained from using either a linear or quadratic objective for weighted

MoMA with objective weights (mutations) on a particular reaction (Figure D.1).

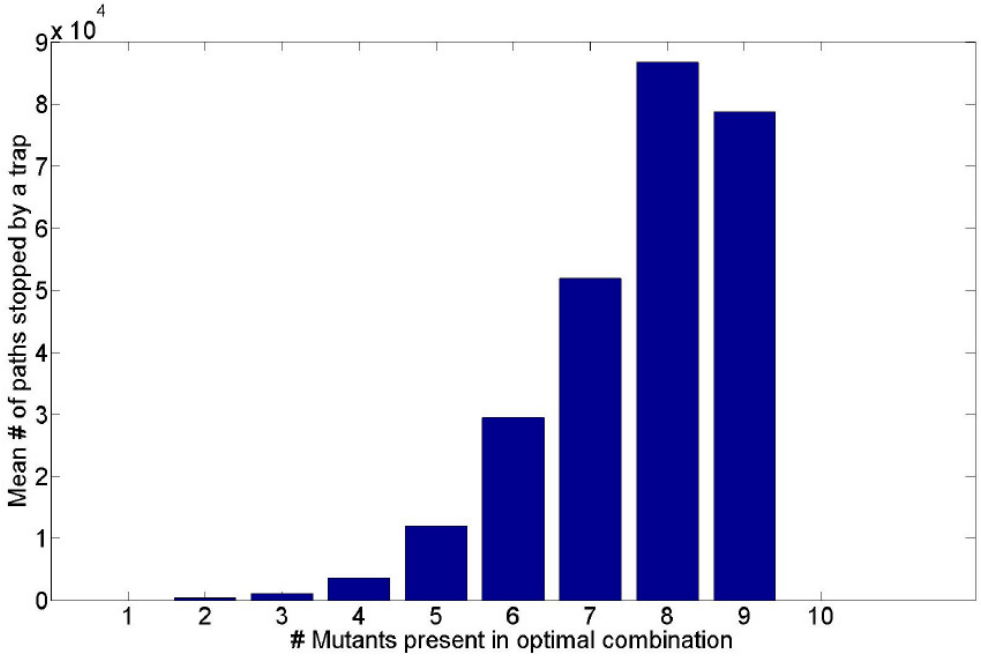
5.2.3 Adaptive trajectories and evolutionary path analysis

Phenotypes involving a small number of mutations have had their adaptive paths analyzed systematically ([173, 174, 177, 178]; four or five mutations). In Weinreich et al. [174], all mutations occur on a single enzyme, and even in such a localized case, sign epistasis, which gives rise to evolutionary traps, occurs for the majority of mutational paths. In the present study, we used less severe grounds for deciding if an evolutionary path is a likely dead-end, but even so, we can see that (Figure 5.2.3; Section D.2.1).

Figure 5.4: An examination of simulated evolutionary dead-ends. Randomly selecting 10 beneficial mutations from over 150 total beneficial mutations in a yeast model will have a varying number of mutations that are present in the combinatorial mutant with the highest fitness The percent of paths that reach the optimal mutant by avoid traps tends to decrease as more mutations are considered (5.5a), or viewed another way, the mean number of paths stopped by traps tends to increase (5.5b).



(a)



(b)

In a genome-scale metabolic model, we observe that while the majority of paths are not likely to have a problem, once eight or nine genes are involved, the average number of paths to reach the optimum without experiencing a trap are only 65%. This

demonstrates that the presence of traps are heterogeneous, just as in the single-enzyme experiment of Weinreich et al. [174], suggesting that the order of mutations often matter to a very significant degree in evolution.

The fact that many unobstructed evolutionary paths are observed at the genome-scale suggests that evolution is, in general, not very predictable (Figure 5.2.3). However, if we consider that neutral mutations are unlikely to fix in a population, the number of viable paths would greatly decrease (Section D.2.1). Interestingly, in every experimental case (Table D.1), only a single fitness peak appears to be present [174, 179]. When we consider simulations from genome-scale models instead of isolated pathways or single protein, we still observe the single fitness optima for five mutation sets, but once we consider eight mutations, every set of eight or more mutations presents with multiple local optima. This suggests that evolution becomes not only less predictable when we look at evolution of larger systems or evolution at a large time scale, but also less likely to reach the global optimal fitness.

Software for evolutionary path analysis

Storing all evolutionary paths in memory or disk can become intractable. For instance, even 10 mutations has $10! \approx 3.6$ million possible paths, but only $2^{10} = 1,024$ mutant combinations. Due to this difficulty, we have created software to allow the dynamic exploration and analyses of these paths.

The C programs for dynamically performing analyses on text files containing fitness data for all combinations of n mutations can be found online, and additional information and and examples related to usage is available (Section D.2.1).

Table 5.1: Pairwise epistasis values (calculated multiplicatively) from two experimental systems.

(a) Khan et al. [178]

fitness		t	s	g	r
1.145	t				
1.108	s	-0.060			
1.030	g	-0.027	-0.014		
1.015	r	-0.056	-0.004	0.005	
1.003	p	0.049	0.078	0.045	0.007

(b) Chou et al. [173]

fitness		gshA	GB	fgh
1.509	gshA			
1.166	GB	-0.120		
1.142	fgh	-0.100	-0.012	
1.096	pntAB	-0.040	0.021	0.029

Small beneficial mutants exhibit positive epistasis

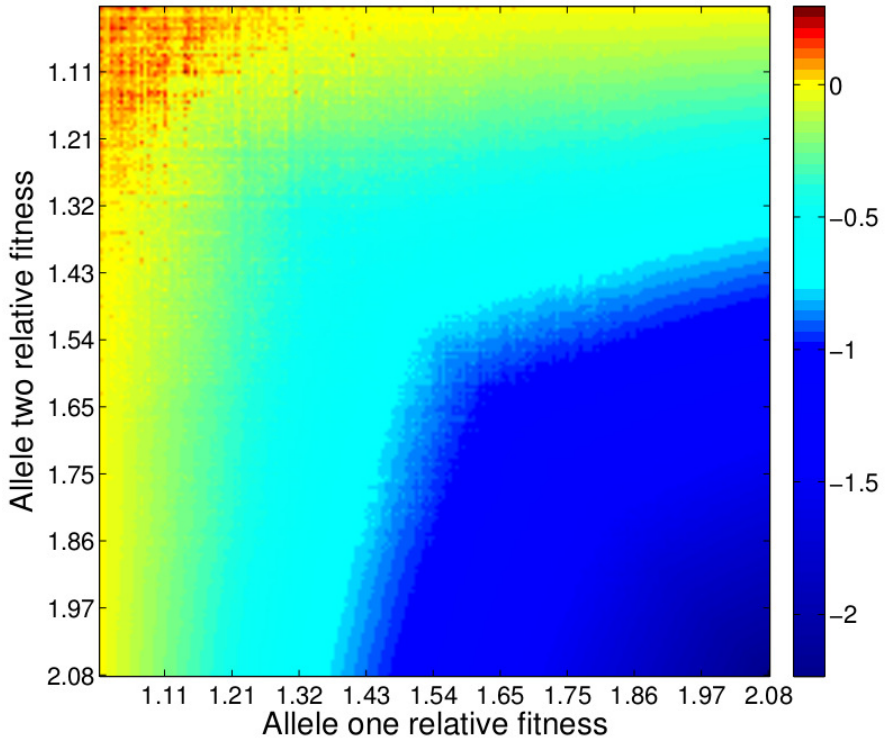
Prior studies using metabolic models have shown that deleterious mutations tend to exhibit negative epistasis for stronger mutations and positive epistasis for weaker mutations ([9, 34]; Figure D.2). The reason for positive epistasis being prevalent for deleterious mutations is that most of them occur in essential pathways, and one mutation will buffer against the effects of a second mutation ([9]). Somewhat surprisingly, we also see that predictions for epistasis involving weakly beneficial mutations also tend to have positive epistasis (Figure 5.6; Section 5.4.1). This trend is verified for two experiments involving multiple genes (as in our simulation) where fitnesses are readily calculated ([173, 178]; Table 5.1).

The trend for increasing negative epistasis as mutations become increasingly beneficial (also termed *diminishing returns epistasis*; [173]) can be easily understood in most contexts, including metabolism, due to the fact that in any given environment there must be a physiological maximum value for most phenotypes, including growth rate of a cell or individual organism. Thus, if one mutant is extremely beneficial, it proportionally limits the effect another beneficial mutant might have when combined together. Such decreasing marginal benefits are not the only reason we may tend to see small mutants with positive epistasis; Fisher's geometric model implies that increasing complexity will

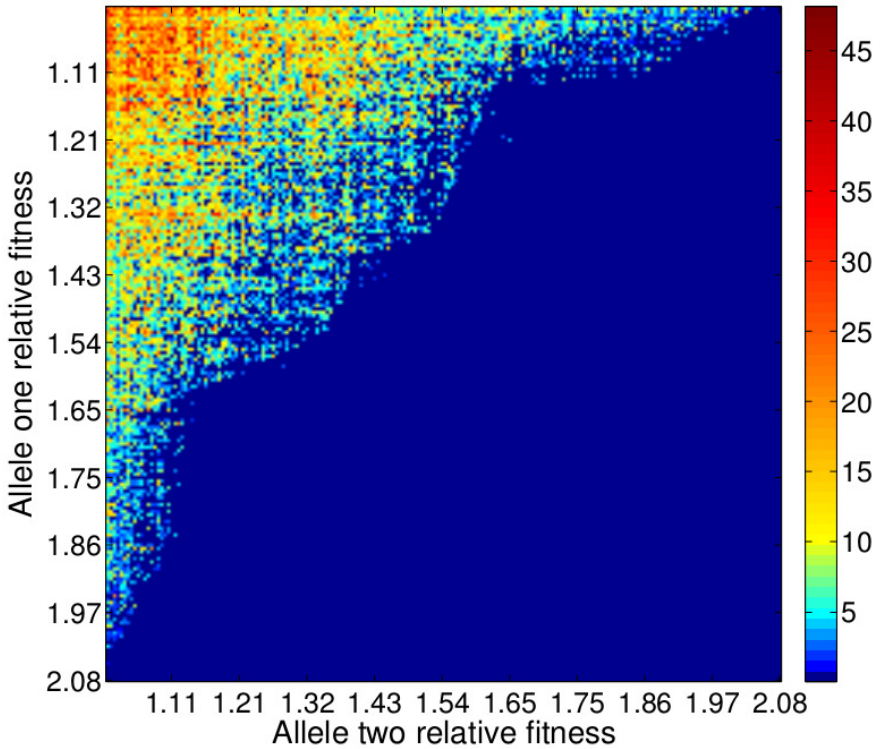
decrease the likelihood that any given mutation will be beneficial, and increasingly so for mutations with more extreme changes in the phenotypic space [180]. Since a mutation that exhibits positive epistasis with many other mutations is just a way of saying the mutation is likely to be beneficial on many genetic backgrounds, the effect observed by Orr [180] extends beyond absolute fitness values and into the epistatic landscape. Furthermore, in our metabolic models (as in most biological systems), fluxes do not operate in isolation; a mutation in a gene expression level or reaction constraint will affect other reaction fluxes in the system, and can be classified as pluripotent. If it is an essential reaction, many fluxes are likely to be affected, thus exhibiting a complex phenotype and a high degree of pluripotency. Large pluripotent mutations would then be expected to have a larger total length in Fisher's geometric model, and be less likely to be beneficial or have positive epistasis. This is exactly what we see in metabolic models with deleterious mutations, which tends to exhibit positive epistasis for smaller mutations and negative epistasis for more extreme mutations (Figure D.2; [9, 34]).

This trend in epistasis can also be captured by our constraint-based modeling framework for beneficial mutations. After *in silico* screening for beneficial mutations (Section 5.4.2), we can examine trends in pairwise epistasis as a function of the single mutant fitnesses Figure 5.6. A fairly distinct border appears to be present between the region that has some positive epistasis and no positive epistasis. By considering that an epistatic cutoff (call it ϵ_c) is employed, we may use the multiplicative epistatic relationship $W_{xy} - W_x W_y > \epsilon_c$, which yields the reciprocal function $W_y < \frac{W_{xy}-\epsilon_c}{W_x}$ for fixed W_{xy}, bounding the region in which positive epistasis, as defined by the threshold ϵ_c, can occur. Since W_{xy} is not a constant, there may be some variation in the actual trend from a true reciprocal function.

The distribution of epistases arising from beneficial mutations is highly similar to



(a) mean epistasis



(b) percent of positive epistasis

Figure 5.6: The mean epistasis (5.6a) and percentage (5.6b) of positive epistasis for epistases such that $|\epsilon| \geq 0.01$.

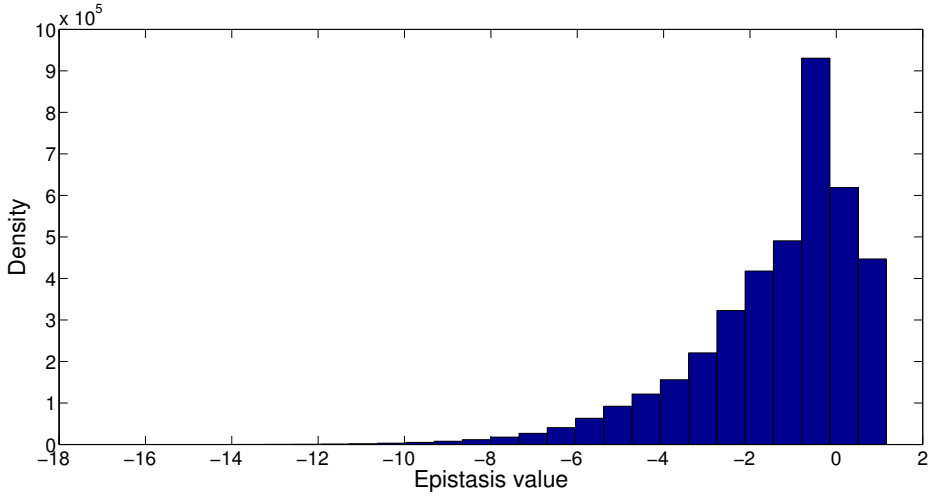


Figure 5.7: Distribution of epistases arising in a yeast model from beneficial mutations sampled according to EVA.

that found in natural biological systems that are also described by population genetic models (Figure 5.7; [181]). Interestingly, the small-scale data from Martin et al. [181] is from an RNA virus (VSV), whereas our stoichiometric models are not designed for modeling viruses. Nonetheless, the trend remains similar, suggesting that this is a trend that extends across completely different types of models as well as from the fitness of complex life forms and simple non-living viruses. This excellent fit requires that an appropriate distribution of fitnesses is sampled according to extreme value theory (EVA; Section 5.4.2; [172, 182]).

5.3 Discussion

We have developed a data-driven modeling framework in the constraint-based metabolic modeling family that allows the exploration of combinatorial epistases and reproduces trends seen from studies in the forefront of evolutionary research [173, 178, 181]. Contemporaneous, theoretical insights in to how a general class of mechanistic models,

encompassing those used in this study, give rise to traditional population genetic models such as Fisher's geometric model [183]. The methodologies described in this paper should have applications in diverse fields due to their reliance on mechanistic models available for many organisms [184].

The questions are often difficult or impossible to assess experimentally due to limited resources. In genome-scale models, to our knowledge, only microbial epistasis has so far been studied for all enzymes (often referred to as genome-scale in this context). This is due to several factors.

One issue is that these computations can still take a significant amount of time, and the increase in model size of Human Recon 2 over Yeast can cause even a relatively simple FBA run to go up by an order of magnitude. This problem is compounded by the increase in the number of genes in the human model, since computing epistasis consumes space and time as $O(n^2)$ where n is the number of genes in the model. More important than this issue, which might be overcome with enough computational resources, is the issue of an objective function. It has been shown numerous times that FBA with a biomass objective can be a reasonable approximation to what a microbe is trying to achieve metabolically [7, 20, 185]. While Recon 2 is equipped with a "generalized biomass reaction", it is not clear what the meaning of this is, and it certainly seems to greatly overestimate the metabolism even of fast-growing cancer cells [186]. We propose FALCON as a potential method to get around this issue for non-microbial models.

Another advantage of FALCON is that it allows one to directly probe mutations that are represented as gene expression perturbations. A decreased level of gene expression may also be metabolically equivalent to the effect of a missense mutation, for example. This allows a different sampling strategy than before; for instance, we could observe

how uniform expression restriction compares to uniform flux restriction [9]. Assuming an accurate model of enzyme-complex expression measurement, the former should be the more realistic model.

A limitation is that we have only considered metabolic genes and their effect on steady-state metabolism. While in principle a similar method could be applied to whole cell models [68, 187], the computation time would not be feasible to the screening for beneficial mutations, nor of exploring them combinatorially, as the time needed for a single mutant takes at least a day even in the smallest bacterial model [68]. Future insights into improving the efficiency of whole cell models, or making a compromise on which systems are simulated (e.g. rFBA, [53]), may improve these efforts.

5.4 Methods

5.4.1 Beneficial mutation simulation for pairwise epistasis

In order to generate a realistic WT flux, we use experimental expression data to fit a flux vector ([1, 188]). Because even FALCON can take one or two orders of magnitude longer than MoMA (or linear MoMA), we use linear MoMA to estimate the flux vectors for single and double mutants. Since our fitness is just flux through the biomass pseudo-reaction, which is a complicated sink, it never seems to carry a flux in practice when expression-flux fitting techniques are used. Therefore, we used experimentally determined growth rates as a constraint in the flux fitting step for calculation of the WT flux vector (strain S96 in YPE (3% ethanol): $\mu = 0.1249$, YPD (2% glucose): $\mu = 0.4621$, strain BY in YPEG: $\mu = 0.21$). Though this approach was only used for pairwise epistasis in the current study, it would also be appropriate for combinatorial epistasis (we

simply used FBA to generate wild-type adapted flux vectors for our combinatorial epistasis simulations).

5.4.2 Mutation screening and sampling

To more closely follow existing theory in population genetics we performed Gaussian sampling centered at the WT phenotype [182]. For each reaction, we sampled flux restrictions where the WT flux (F_{WT}) was at the mean of the Gaussian distribution. Since there is some uncertainty in the underlying distribution of mutational effects that would best reflect nature, we used several different distributions and employed the FVA to bound the distribution [37]. For example, for the 99.9% Gaussian sampling (Figures 5.7 and 5.8), for each reaction, we sampled fluxes from a truncated normal distribution between the FVA minimum ($F_{\min}$) and FVA maximum ($F_{\max}$) where the underlying normal distribution would have 99.9% of its density in $F_{WT} \pm \max\left(\left|F_{WT} - F_{\min}\right|, \left|F_{WT} - F_{\min}\right|\right)$. This sampling technique (in particular the symmetric bounds about F_{WT}) was chosen in order to approximate how sampling mutations from Fisher's geometric model results in a beneficial mutation distribution distribution conforming to extreme value distributions [182, 189]. Only beneficial mutants were kept for the present analysis.

A truncated normal sampling strategy drawing 99.9% of mutations within the larger FVA bound will be less likely to sample extremely divergent fluxes than a sampling strategy drawing 50% of mutations within the larger FVA bound since the tails of the latter distribution will be placed further away from the WT. We find it is preferable to sample a high percentage of mutations that are closer to the WT in order to generate beneficial mutations distribution conforming to extreme value theory, since otherwise the mutation sampling in our metabolic model begins to approximate a uniform distribution

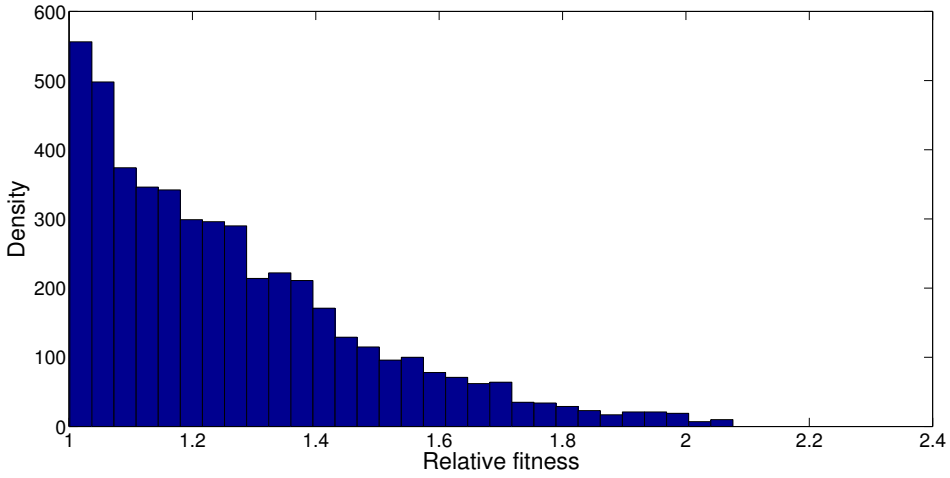


Figure 5.8: Distribution of beneficial mutations arising in a yeast YPE model sampled according to EVA (truncated normal sampling with 99.9% mutations occurring within the larger FVA bound).

(Figure D.3). It can be seen that epistasis distributions resulting from EVA-conforming fitness distributions result in a skew toward near-zero epistasis and positive epistasis as we tend to sample more mutants near the WT (Figure 5.7).

5.5 Acknowledgments

We thank Xiaoxian Guo and Zhe Wang for measuring the growth rate of yeast in YPE and YPEG media, respectively, as well as discussing yeast biology and experiments more generally.

APPENDIX A

SUPPORTING INFORMATION FOR DYNAMIC EPISTASIS FOR DIFFERENT ALLELES OF THE SAME GENE

A.1 Supporting Figures

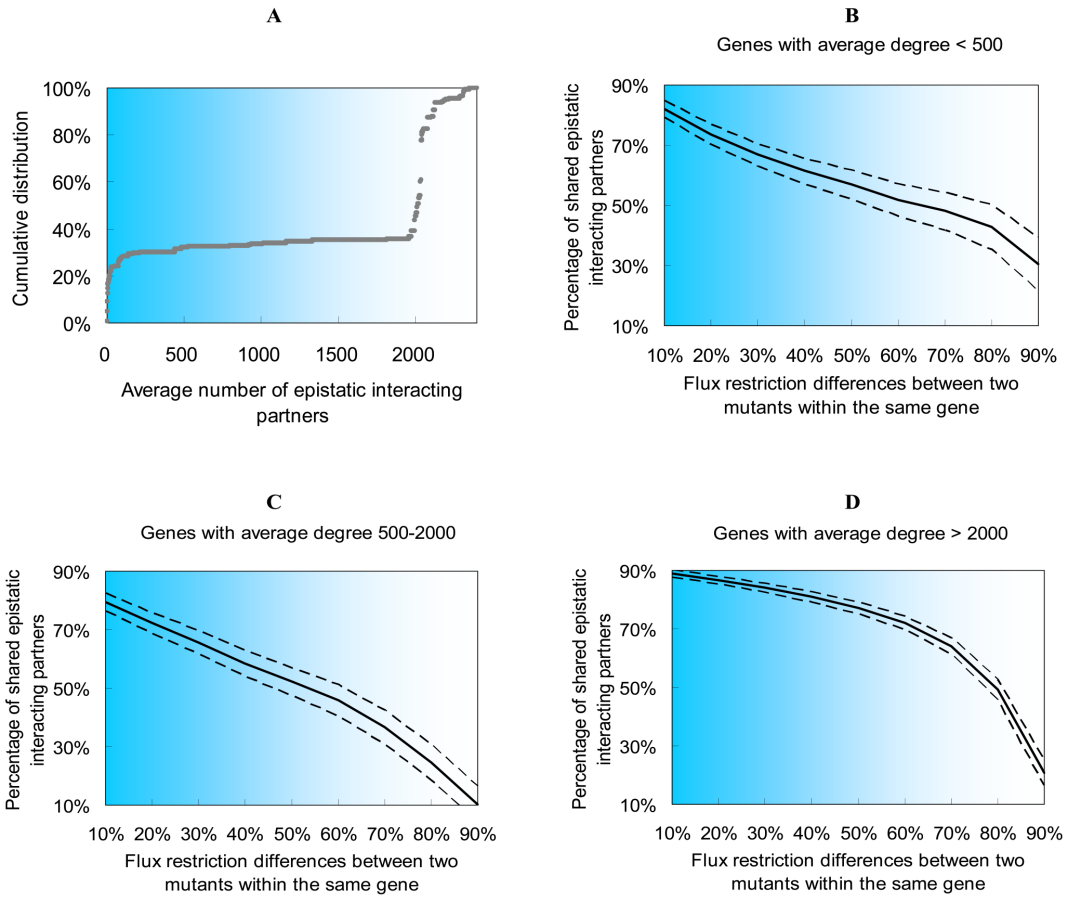


Figure A.1: The conclusion in Fig. 1 is not dependent on the average number of epistatic interaction partners per gene. (A) The distribution of average number of epistatic interaction partners per gene. For each gene with epistasis, its average number of epistatic interaction partners was calculated among all mutant alleles of this gene. (B-D) A similar conclusion to that of Fig. 1 can be obtained when we only use genes with fewer than 500 (B), 500-2,000 (C), and more than 2,000 (D) average epistatic interaction partners. The same methods in Fig. 1 were used here to generate B-D.

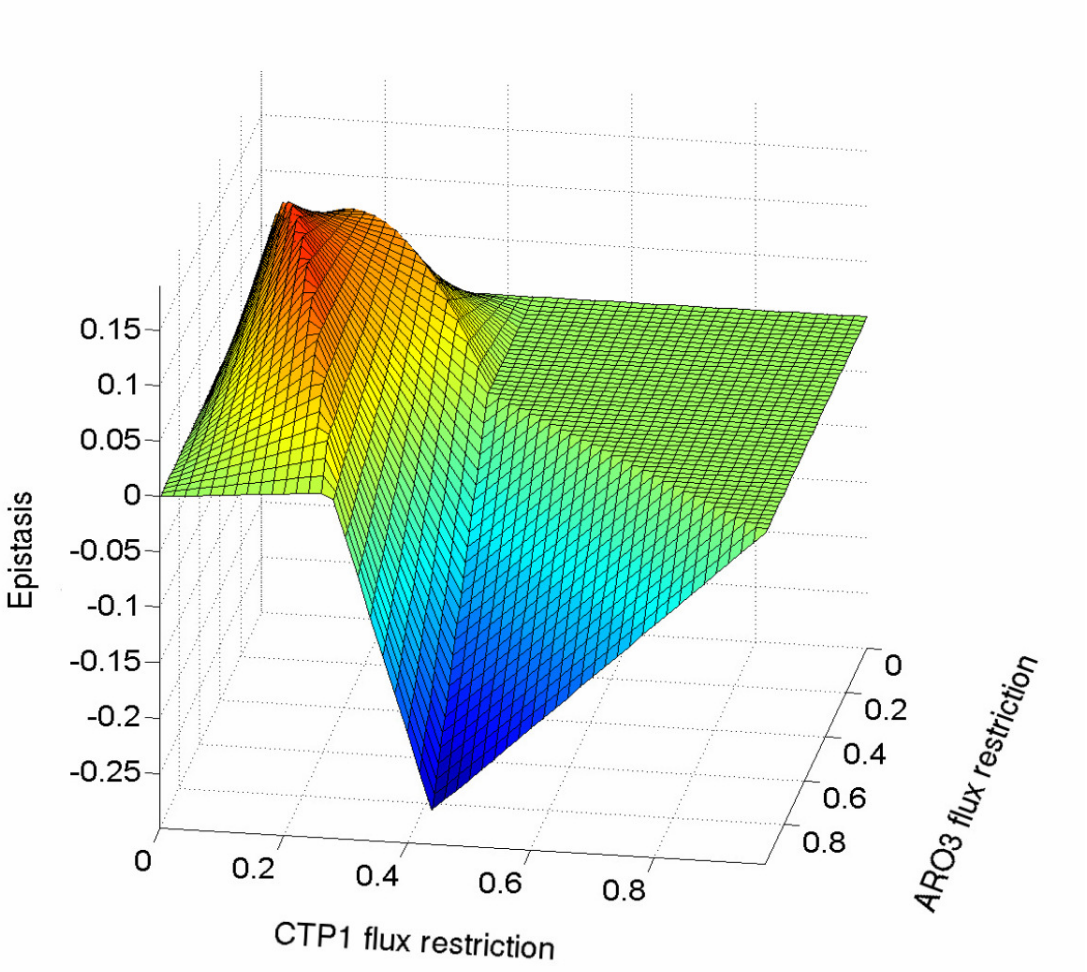


Figure A.2: A complex epistatic landscape exhibits a transition from large positive to large negative epistasis values, along with a region of zero epistasis. Epistasis is viewed as a function of the CTP1 and ARO3 genes flux restriction. The color corresponds to the z-axis (epistasis), with red being more positive, green being near zero, and blue being more negative.

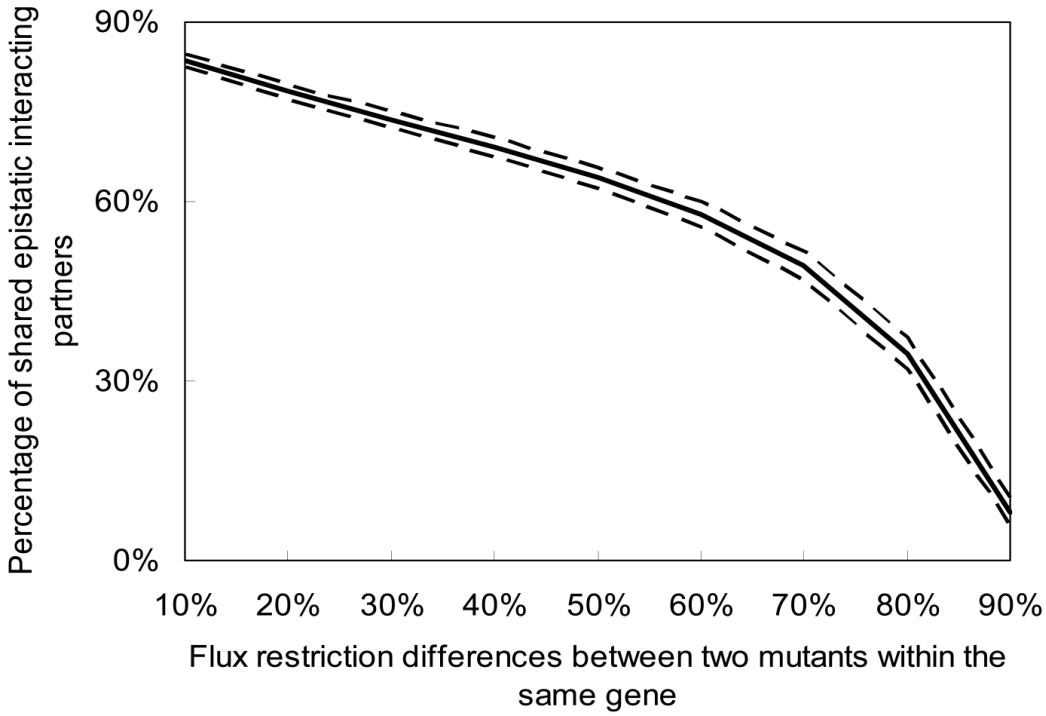
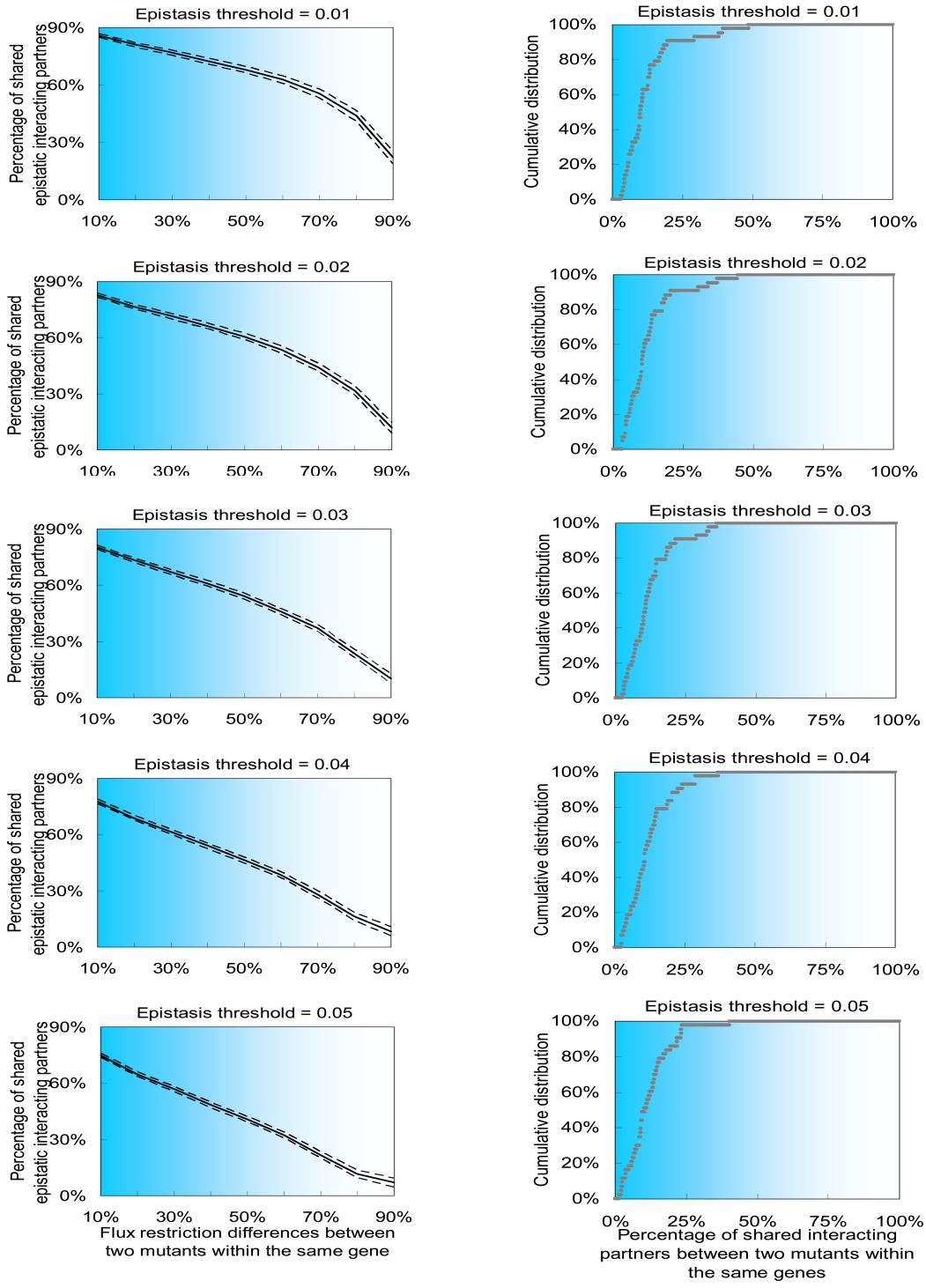


Figure A.3: Percentage of shared epistatic interacting partners based on flux differences between two mutant alleles of the same gene. The analysis procedure is the same as Fig. 1A, but instead of using the S. cerevisiae model, here we repeated the analysis using the E. coli model (38).

Figure A.4: The conclusion that epistatic relations between genes are allele-specific is robust to various epistasis thresholds. Left 5 panels: The FBA simulation results for the distribution of the percentage of shared epistatic interaction partners between two mutant alleles within the same gene. Solid and broken lines represent mean and 95% confidence intervals, respectively. Right 5 panels: The cumulative distribution for the percentage of shared epistatic interaction partners between two mutant alleles within the same gene based on real experimental data. Both experimental and simulated results are robust under various epistasis thresholds.


Epistasis threshold = 0.01
Percentage of shared epistatic interacting partners
90%
60%
30%
0%
10%
30%
50%
70%
90%
Cumulative distribution
100%
80%
60%
40%
20%
0%
0%
25%
50%
75%
100%
Epistasis threshold = 0.02
Epistasis threshold = 0.03
Epistasis threshold = 0.04
Epistasis threshold = 0.05
Flux restriction differences between two mutants within the same gene
Percentage of shared interacting partners between two mutants within the same genes

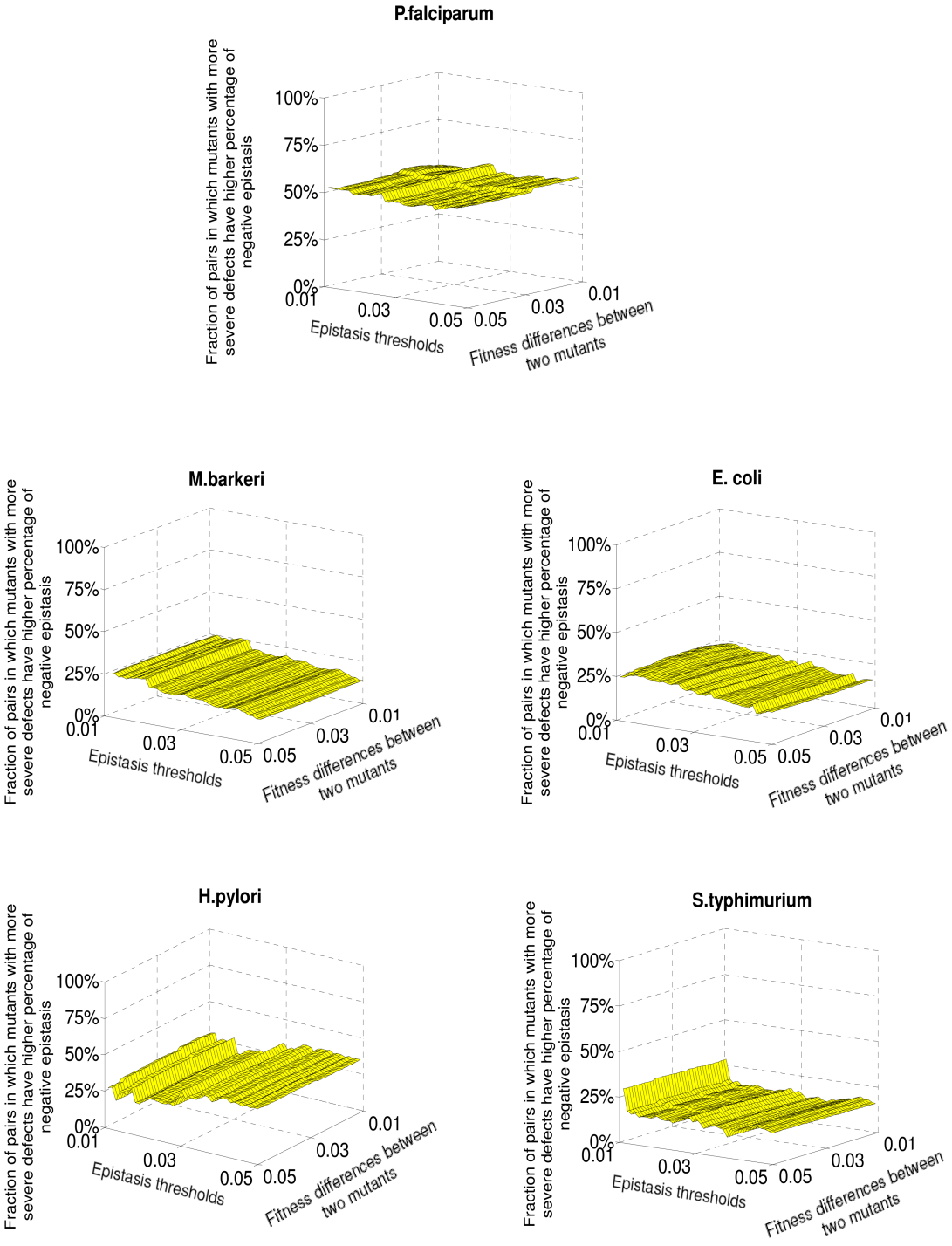


Figure A.5: The conclusion in Fig. 3, for which mutant alleles with more severe defects tend to have a higher percentage of negative epistasis in eukaryotes than bacteria and archaea, is robust under various epistasis and fitness difference thresholds. The same methods to generate Fig. 2C for S. cerevisiae are used here for the other five species.

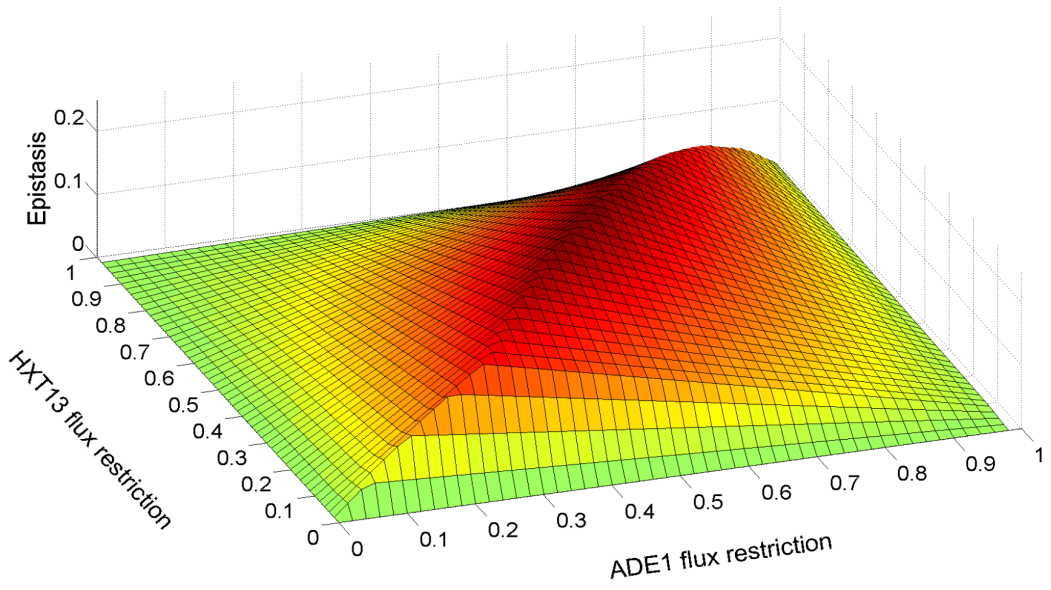


Figure A.6: An epistatic landscape exhibits smooth change in epistasis as a function of the flux restriction for the genes HXT13 and ADE1. The color corresponds to the z-axis (epistasis), with red being more positive, and green being near zero. See dataset S3 for simulated data. HXT13 is a hexose transporter and ADE1 is required for de novo purine biosynthesis. The epistasis surface for HXT13 and ADE1 is quite smooth, which is a fairly common pattern and we may infer that epistasis, at least in metabolism, is often dependent on thresholds.

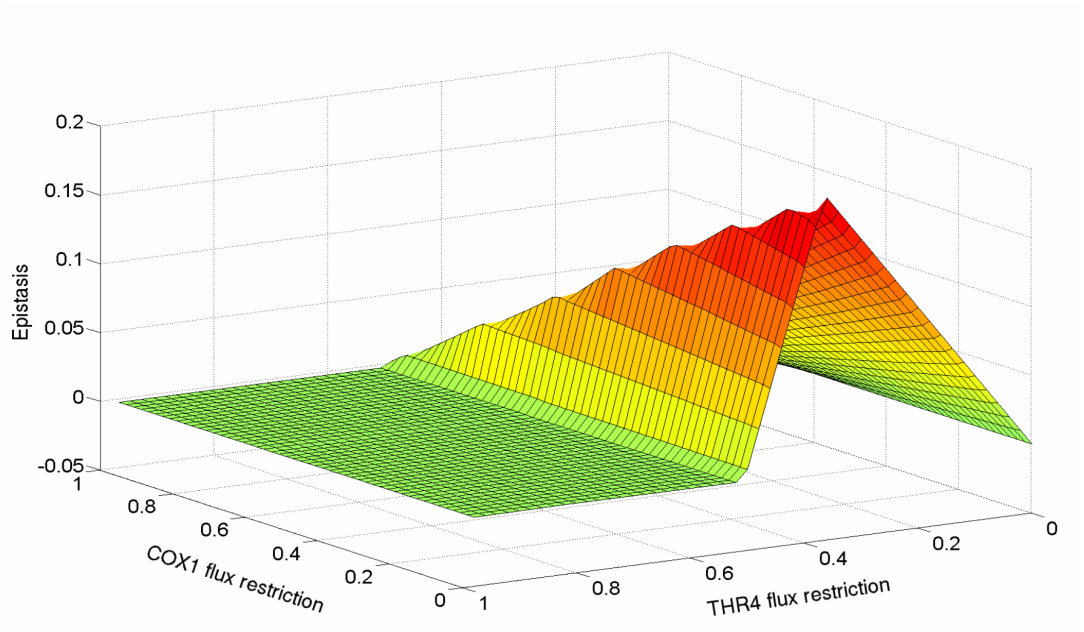


Figure A.7: An epistatic landscape exhibits a sharp transition to zero epistasis, primarily as a consequence of the THR4 flux restriction. The color corresponds to the z-axis (epistasis), with red being more positive, and green being near zero. See dataset S4 for simulated data. Epistasis is examined between threonine synthase gene THR4 and COX1 (subunit 1 of cytochrome c oxidase). Both genes are associated with mutually exclusive reactions. As shown in the figure, there are regions where the epistasis is effectively zero (on the order of 10-5) where the THR4 single mutant growth rate has only changed very slightly, effectively allowing the mutations to act independently. Once the THR4 mutant becomes more severe, the effects are no longer independent.

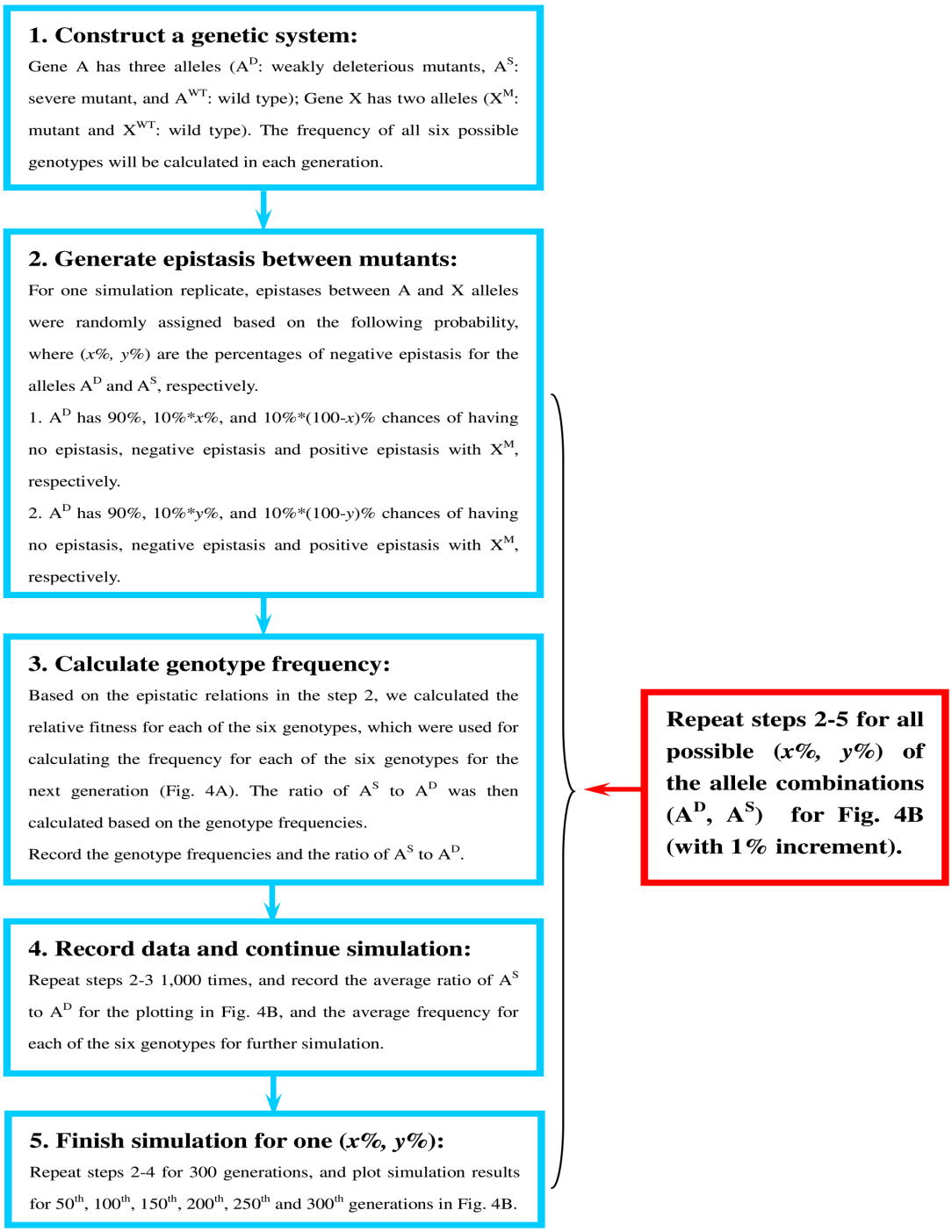


Figure A.8: A flow chart to illustrate the simulation process that generates Fig. 4B. This procedure included 5 steps as indicated in the 5 blue boxes, and we have repeated step 2 to step 5 in the simulation to produce all possible allele combinations, as highlighted in the red box.

A.2 Supporting Data

Supporting datasets S1-S4 are available online (DOI: 10.1073/pnas.1121507109).

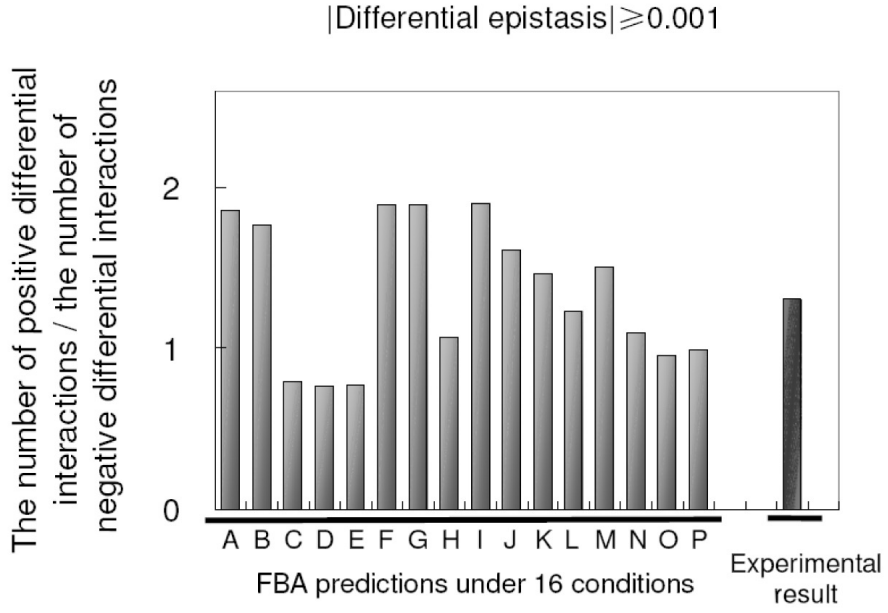
APPENDIX B

SUPPORTING INFORMATION FOR DYNAMIC EPISTASIS UNDER VARYING ENVIRONMENTAL PERTURBATIONS

B.1 Supporting Figures

Figure B.1: More positive differential epistases under environmental perturbations for different thresholds of differential epistasis ($|d\epsilon| \geq 0.001$, **A**) and ($|d\epsilon| \geq 0.05$, **B**). Ratio of positive to negative differential epistases in each simulated condition are shown. The letters A-P represent acetaldehyde, acetate, adenosine 3',5'-bisphosphate, adenosyl methionine, adenosine, alanine, allantoin, arginine, ethanol, glutamate, glutamine, glycerol, low glucose, phosphate, trehalose, and xanthosine, respectively.

A


|Differential epistasis|⩾0.001
The number of positive differential interactions / the number of negative differential interactions
0
1
2
A B C D E F G H I J K L M N O P
FBA predictions under 16 conditions
Experimental result

B

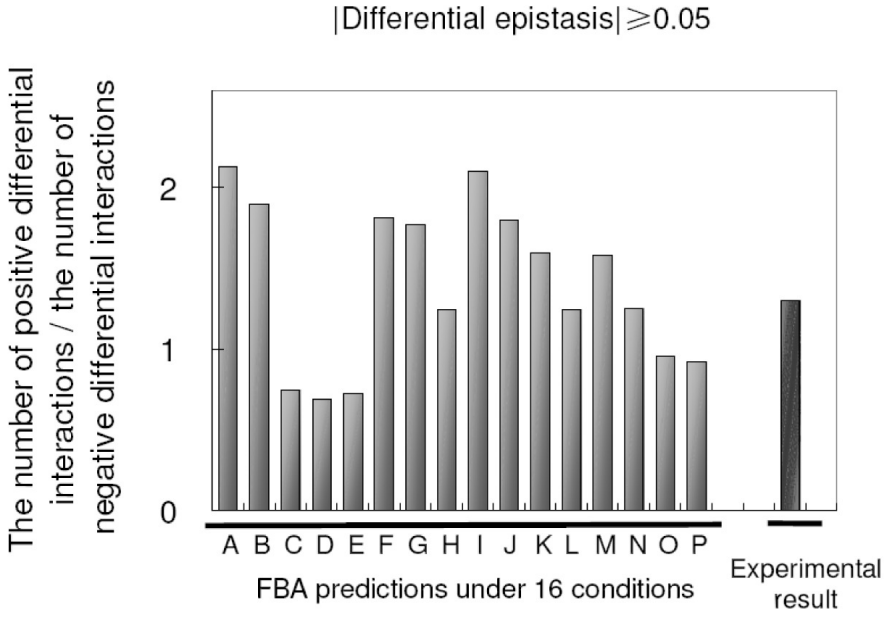
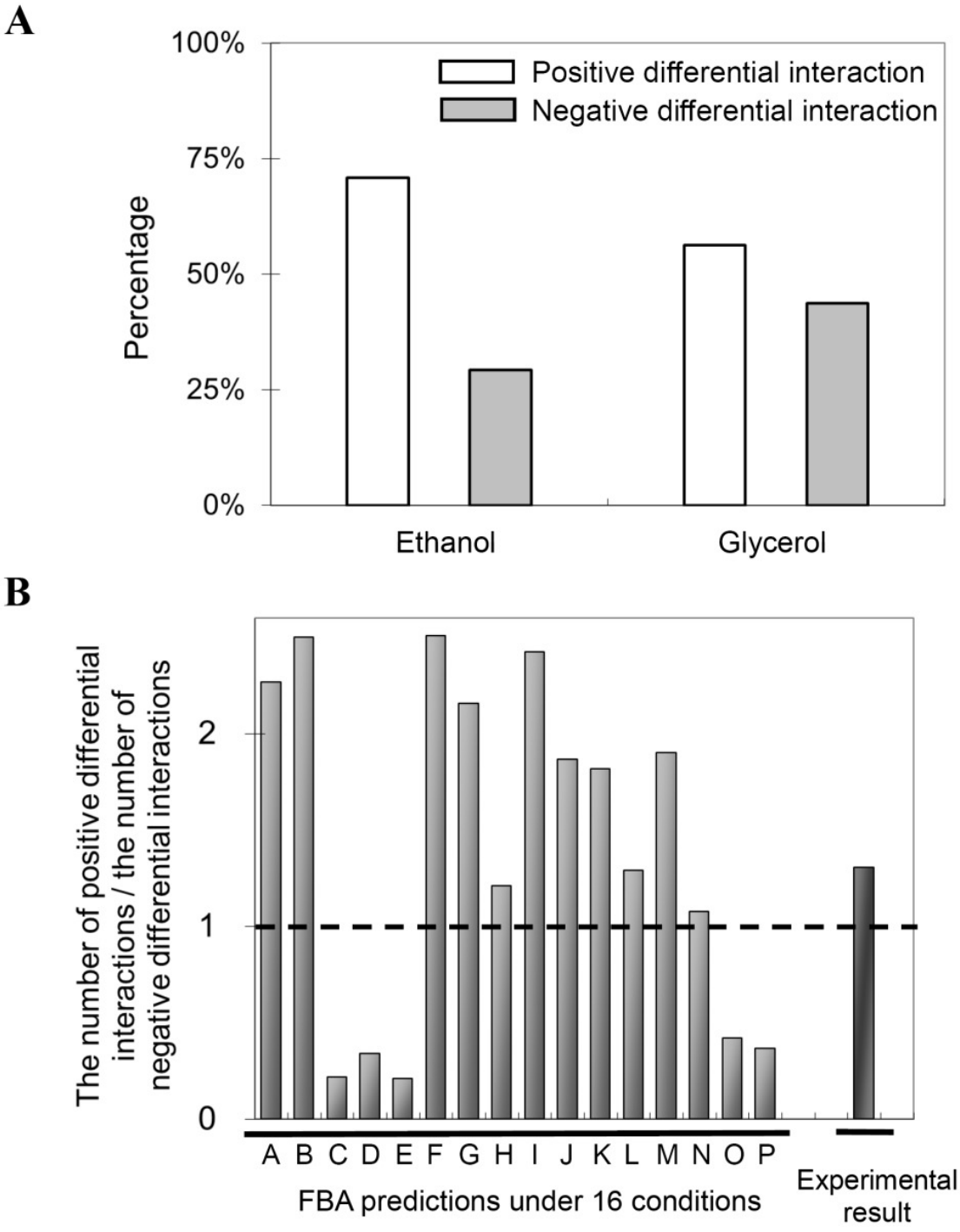

|Differential epistasis|⩾0.05
The number of positive differential interactions / the number of negative differential interactions
0
1
2
A B C D E F G H I J K L M N O P
FBA predictions under 16 conditions
Experimental result

Figure B.2: Analogous to Figure 3.3.2B-C, but using a maximum growth rate for each condition, where the maximum is constrained to be no higher than the high-glucose growth rate. (**A**) Percentage of positive and negative differential epistases under ethanol and glycerol conditions. (**B**) Ratio of positive to negative differential epistases in each simulated condition. The result from a high-throughput experiment is also shown. The letters A-P represent acetaldehyde, acetate, adenosine 3',5'-bisphosphate, adenosyl methionine, adenosine, alanine, allantoin, arginine, ethanol, glutamate, glutamine, glycerol, low glucose, phosphate, trehalose, and xanthosine, respectively. Note that in (B), low glucose has the same growth rate as high-glucose, but has different epistatic interactions since we still use the high-oxygen uptake level associated with the low glucose condition.


A
100%
75%
50%
25%
0%
Percentage
Positive differential interaction
Negative differential interaction
Ethanol
Glycerol
B
The number of positive differential interactions / the number of negative differential interactions
2
1
0
A B C D E F G H I J K L M N O P
FBA predictions under 16 conditions
Experimental result

A

Epistasis threshold = 0.001

		Ethanol		
		Positive Epistasis	Negative Epistasis	No Epistasis
Glycerol	Positive Epistasis	27,494	175	3,537
	Negative Epistasis	818	205	783
	No Epistasis	2,240	104	373,704

Epistasis threshold = 0.05

		Ethanol		
		Positive Epistasis	Negative Epistasis	No Epistasis
Glycerol	Positive Epistasis	22,095	1	264
	Negative Epistasis	633	13	493
	No Epistasis	3,105	194	382,262

B

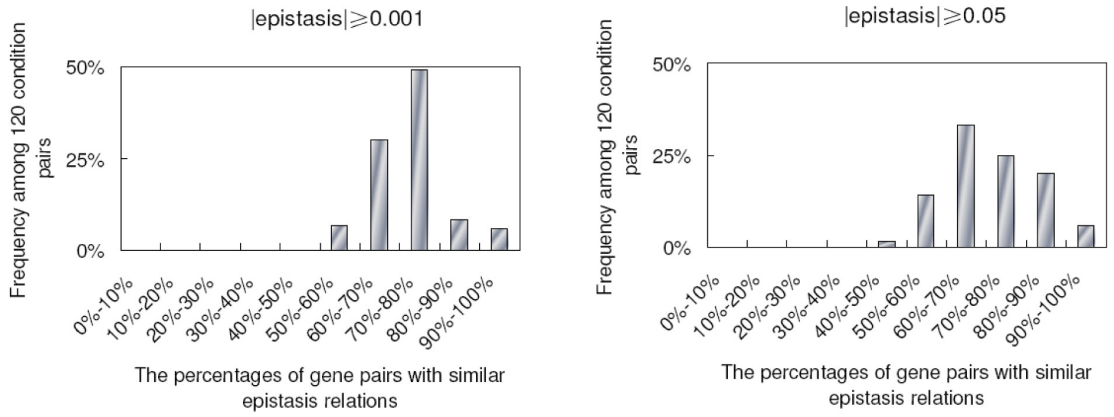


Figure B.3: Epistasis dynamics between environmental perturbations under different epistasis definition. (**A**) Number of gene pairs with various epistatic relationships between ethanol and glycerol growth conditions under a lower ($|\epsilon| \geq 0.001$) and a higher ($|\epsilon| \geq 0.05$) epistasis threshold. (**B**) The distribution for the percentages of gene pairs with similar epistasis relations between any 2 of 16 conditions under a lower ($|\epsilon| \geq 0.001$) and a higher ($|\epsilon| \geq 0.05$) epistasis threshold.

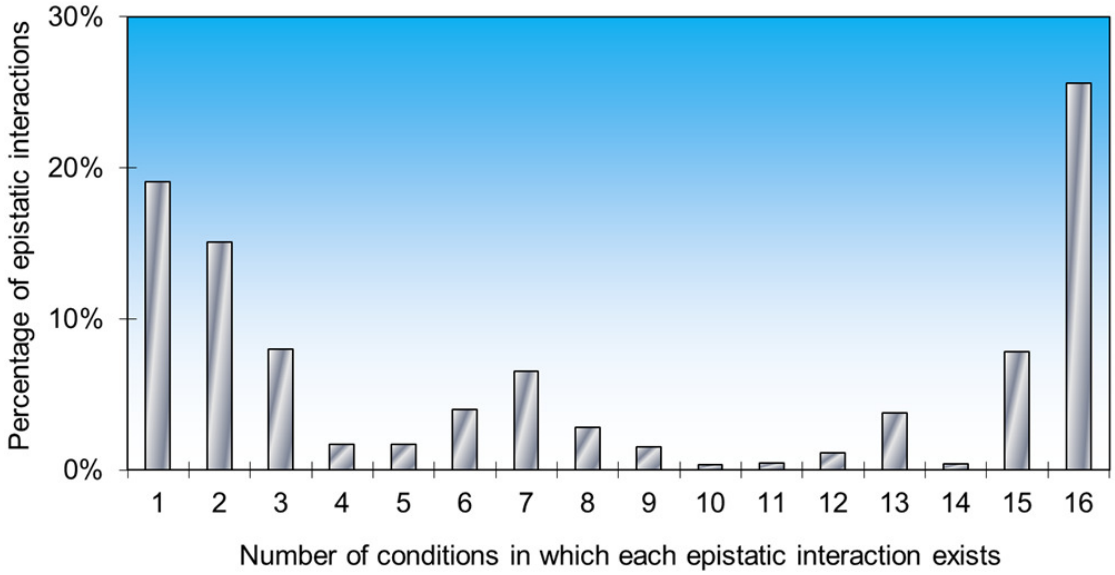


Figure B.4: Analogous to Figure 3.3.3A, but using a maximum growth rate for each condition, where the maximum is constrained to be no higher than the high-glucose growth rate. Distribution for the number of conditions in which each epistatic interaction exists. Note that $\approx 26\%$ of epistatic relations are extremely stable (the very right bar) and $\approx 19\%$ are extremely dynamic (the very left bar).

B.2 Supporting Tables

Tables S1-S6 are available online at `https://app.box.com/s/x1bx3bntyzlpbh3ciuwq`.

B.2.1 Table Legends

Table S1. Wild-type growth rates used in the maximal growth rate simulations used for Figures B.1 and B.4.

Table S2. Condition-specific epistases and sign-epistases prevalence in the iMM904

yeast model.

Table S3. GO term enrichment analysis results for differential epistasis in transition to ethanol.

Table S4. Properties of simulated systems that correlate with the ratio of positive to negative differential epistases.

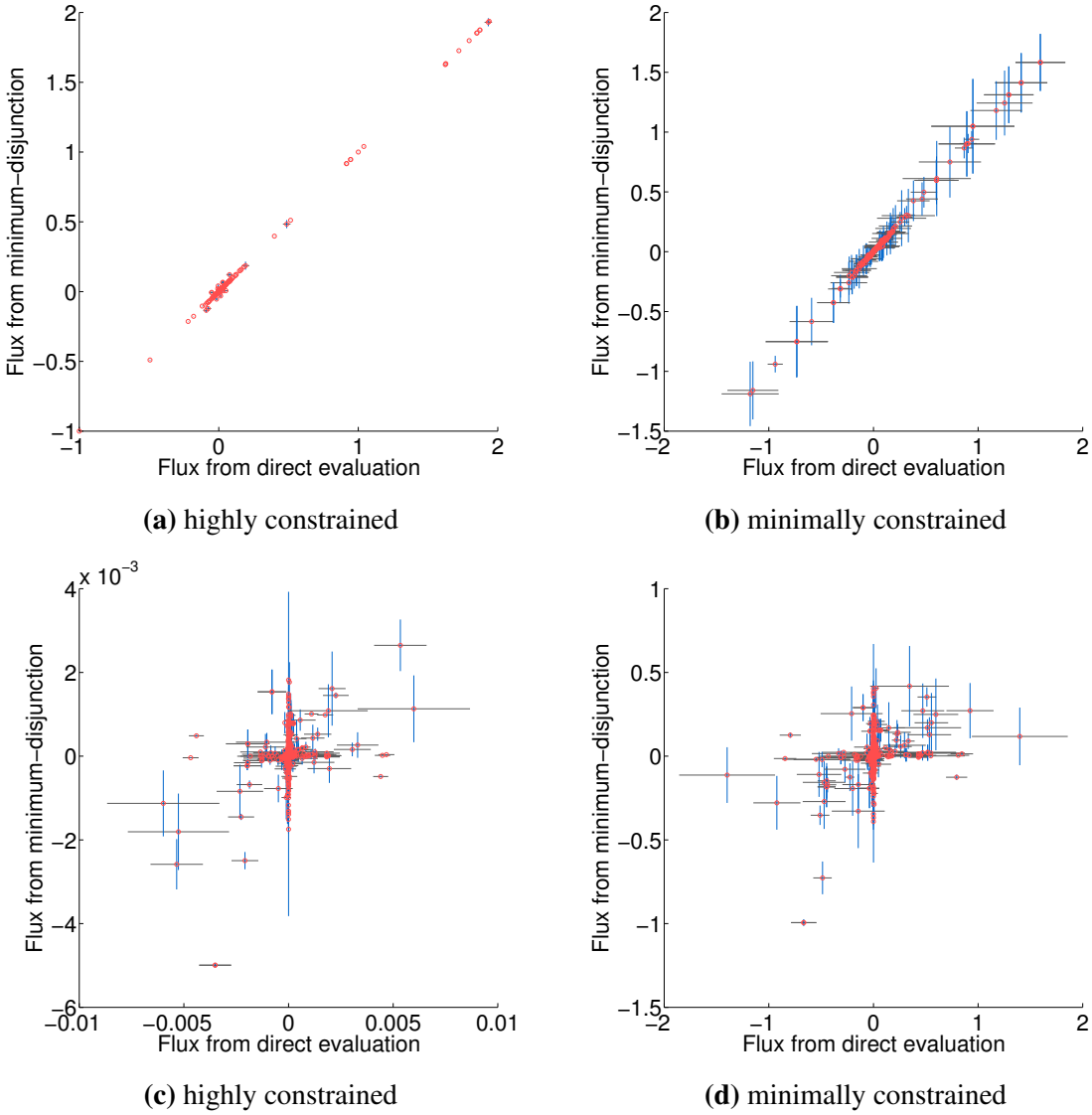
Table S5. List of epistatic interactions for the extremely stable, dynamic, and intermediate epistasis networks.

Table S6. Table of network parameters for stable, dynamic, and intermediate epistasis.

APPENDIX C

FALCON

C.1 Supporting figures



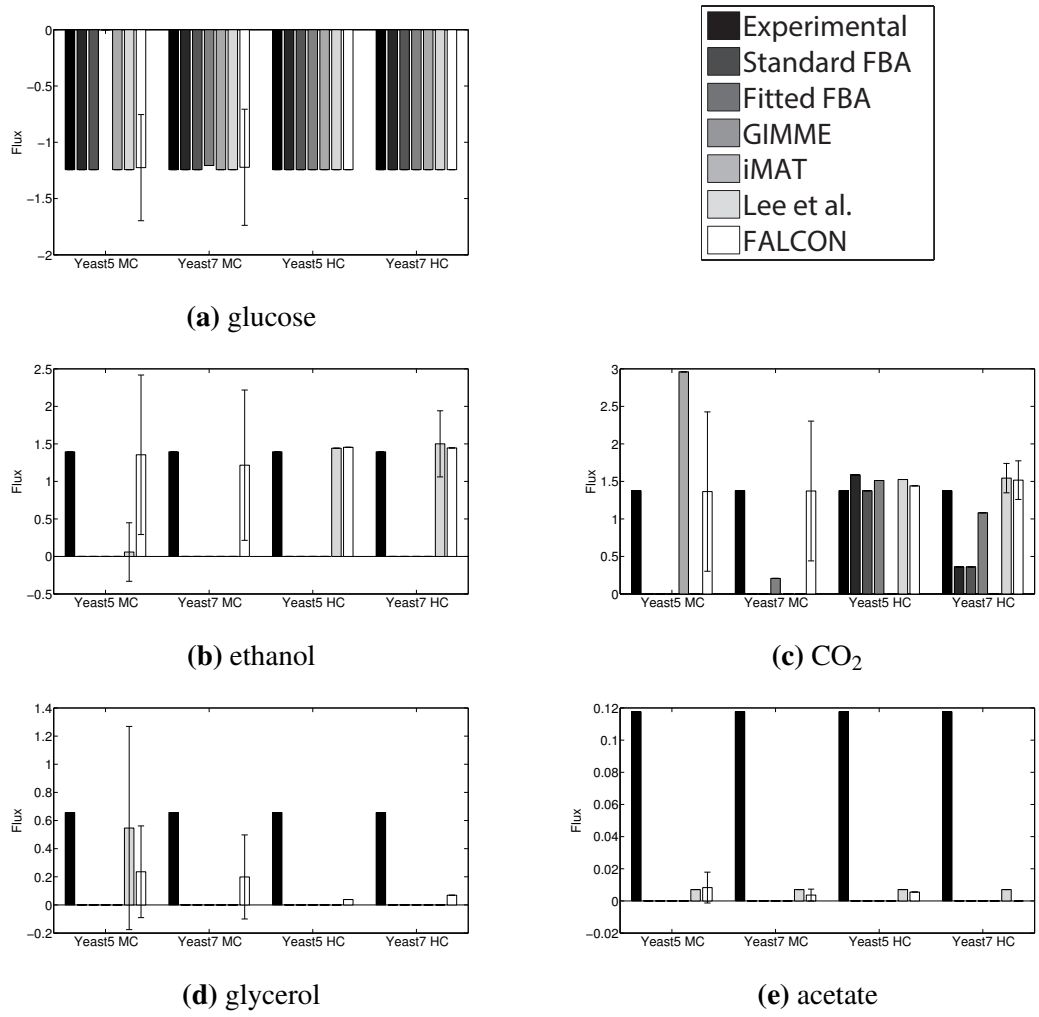
(a) highly constrained

(b) minimally constrained

(c) highly constrained

(d) minimally constrained

Figure C.1: Comparison of fluxes when FALCON is run with enzyme abundance calculated by direct evaluation (x-axis) and the minimum disjunction algorithm (y-axis); error bars with length equal to one standard deviation are shown for both approaches as a result of alternative solutions in FALCON. Yeast was evaluated with default (highly) constrained **(a)** and minimally constrained **(b)** models, and no strong difference between direct evaluation or or the minimum disjunction method is observed in either case. However, for human models with a highly constrained reaction set (RPMI media, CORE-sign, and enzymatic direction) **(c)** and default constraints **(d)**, we see there is a large amount of variation between the two evaluation techniques. In the human cases, two outliers were not shown that correspond to a single large flux cycle ('release of B12 by simple diffusion' and 'transport of Adenosylcobalamin into the intestine').



(a) glucose

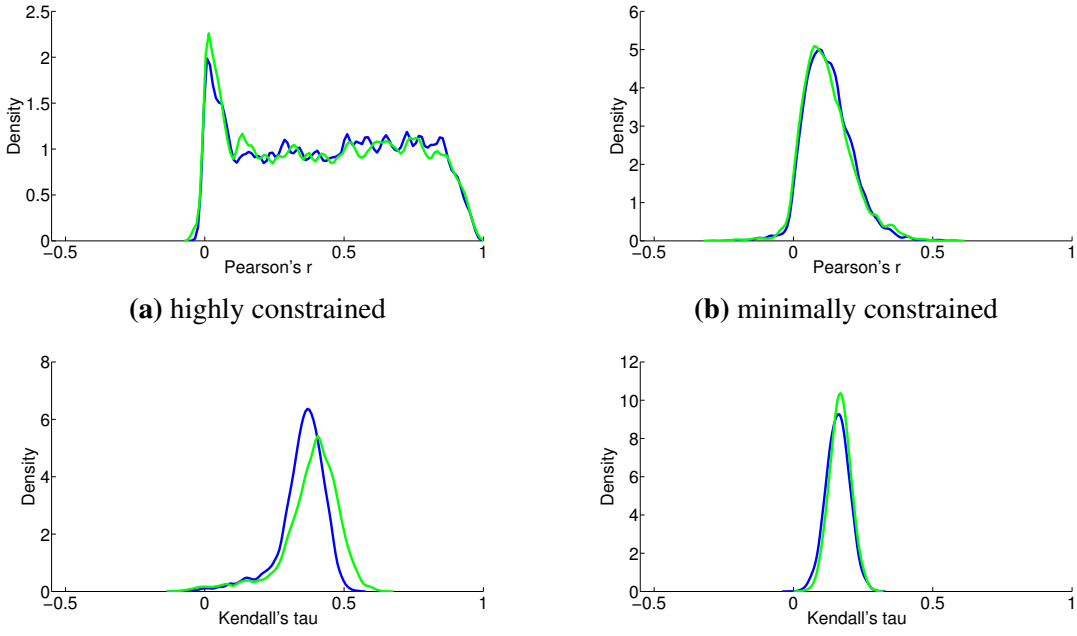
(b) ethanol

(c) CO_2

(d) glycerol

(e) acetate

Figure C.2: Shown are flux predictions using a number of methods and four different models (Yeast 5 MC and Yeast 7 MC are minimally constrained Yeast 5 and Yeast 7; Yeast 5 HC and Yeast 7 HC are highly constrained Yeast 5 and Yeast 7). Error bars are shown for the Lee et al. method and for FALCON, where one side of the error bar corresponds to a standard deviation. Note that there can be no variation for glucose in the former case since glucose flux is fixed as part of the method. FALCON performs very well for large fluxes **(a-c)**, and is also the best performer in general for the next largest flux, glycerol **(d)**. It also has sporadic success for smaller fluxes, but all methods seem to have trouble with the smallest fluxes (e.g. **e**). Note that fluxes are drawn in log scale (specifically a flux v is drawn as $\mathrm{sgn}(v)\log_{10}(1+v)$). Similar results are obtainable for the 85% maximum growth condition.



(a) highly constrained

(b) minimally constrained

(c) highly constrained

(d) minimally constrained

Figure C.3: Kernel-smoothed PDFs are drawn for correlations between the entire flux vector estimated by FALCON on permuted and unpermuted data. Stability and correlation are effected by constraints, as there are differences between the minimally constrained **(b, d)** and highly constrained **(a, c)** Yeast 7 models. 5,000 permutation replicates were performed in all cases.

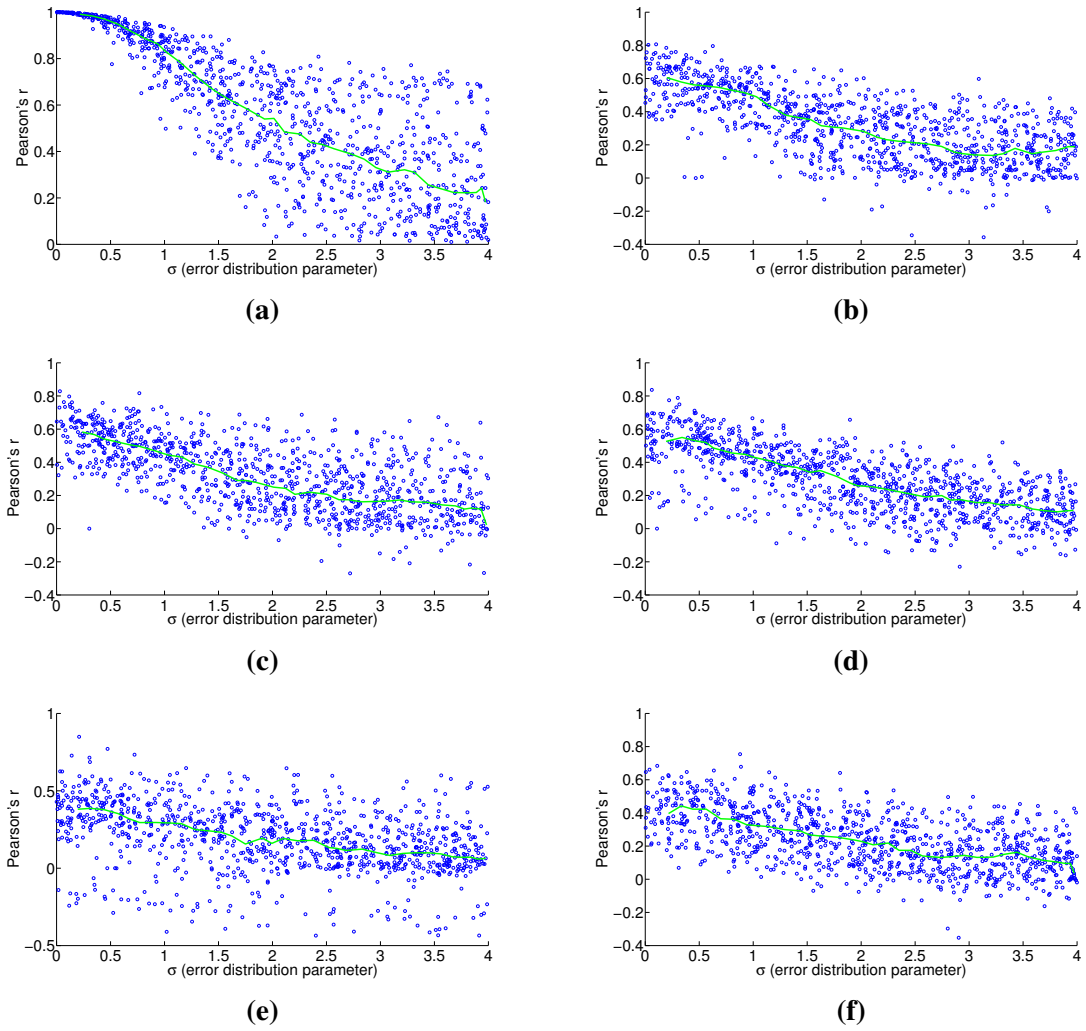
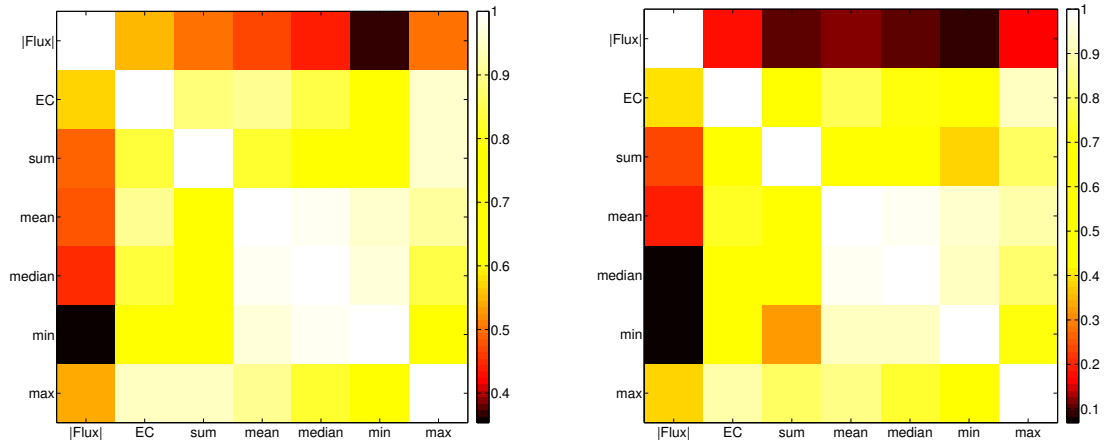


Figure C.4: These figures are generated in the same way as those in Figure 4.4.2, but for Human Recon 2 instead of Yeast 7. We used several different constraint sets based on experimental media and exometabolic flux data in the NCI-60 cell lines [3]. These constraints were applied cumulatively, and are listed in the order of most constrained **(b)** to least constrained **(f)**. Included are default Recon 2 constraints **(f)**, RPMI media constraints (**e**; function `constrainCoReMinMaxSign`; 556 constraints), exometabolic fluxes with a common sign across all cell lines and replicates (**d**; function `constrainCoReMinMaxSign`; 567 cumulative constraints), enzymatic reaction directionality constraints from a linear MoMA fitting on the exometabolic flux data that agree across all NCI-60 cell lines (**c**; `constrainImputedInternal`; 593 cumulative constraints), and the same again considering all reactions instead of only enzymatic reactions (**b**; 618 cumulative constraints).



(a) yeast **(b)** human

Figure C.5: Pearson correlation between FALCON flux magnitudes, prerequisite enzyme complex estimates (from minDisj), and various simpler gene expression estimates based on the list of genes associated to each reaction. For yeast **(a)**, the upper and lower triangles are the 75% and 85% maximum growth conditions, respectively, and human is done similarly with the K562 and MDA-MB-231 cell lines **(b)**. As for expression estimates, the sum of expression and enzyme complex estimate levels are generally the least correlated with other expression estimates. As expected, the enzyme complex estimates are the most correlated with the FALCON fluxes, as they are used in the algorithm. However, it is important to note that they are not very similar, exemplifying the affect the network constraints play when determining flux. Interestingly, enzyme complex abundance is found to correlate very highly with the maximum expression level for the complex; this can be attributed to many genes having relatively simple complexes that are isozymes, where one major isozyme is typically highly expressed.

C.2 Assumptions for enzyme complex formation

In order to quantify enzyme complex formation (sometimes called enzyme complexation), the notion of an enzyme complex should be formalized. A protein complex typically refers to two or more physically associated polypeptide chains, which is sometimes called a quaternary structure. Since we are not exclusively dealing with multiprotein complexes, we refer to an enzyme complex as being one or more polypeptide chains that act together to carry out metabolic catalysis.

Assumption 1. A fundamental assumption that we need in order to guarantee an accurate estimate of (unitless) enzyme complex abundance are the availability of accurate measurements of their component subunits. Unfortunately, this is currently not possible, and we almost always must make do with mRNA measurements, which may even have some degree of inaccuracy in measuring the mRNA abundance. What has been seen is that Spearman's $\rho = 0.6$ for correlation between RNA-Seq and protein intensity in datasets from HeLa cells [190]. This implies that much can likely still be gleamed from analyzing RNA-Seq data, but, an appropriate degree of caution must be used in interpreting results based on RNA-Seq data. By incorporating more information, such as metabolic constraints, we hope to obviate some of the error in estimating protein intensity from RNA-Seq data.

Assumption 2. We also include the notion of isozymes–different proteins that catalyze the same reaction–in our notion of enzyme complex. Isozymes may arise by having one or more differing protein isoforms, and even though these isoforms may not be present in the same complex at the same moment, we consider them to be part of the enzyme complex since one could be substituted for the other.

As an example for assumptions described so far, take the F_1 subcomplex of ATP Synthase (Figure C.6), which is composed of seven protein subunits (distinguished by color, left). On the right-hand side we see different isoforms depicted as different colors. Error in expression data aside, instead of considering the abundances with multiplicity and dividing their expression values by their multiplicity, it may be easier to simply note that the axle peptide (shown in red in the center of the complex) only has one copy in the complex, so its expression should be an overall good estimation of the F_1 subcomplex abundance. This reasoning will be useful later in considering why GPR rules may be largely adequate for estimating the abundance of most enzyme complexes.

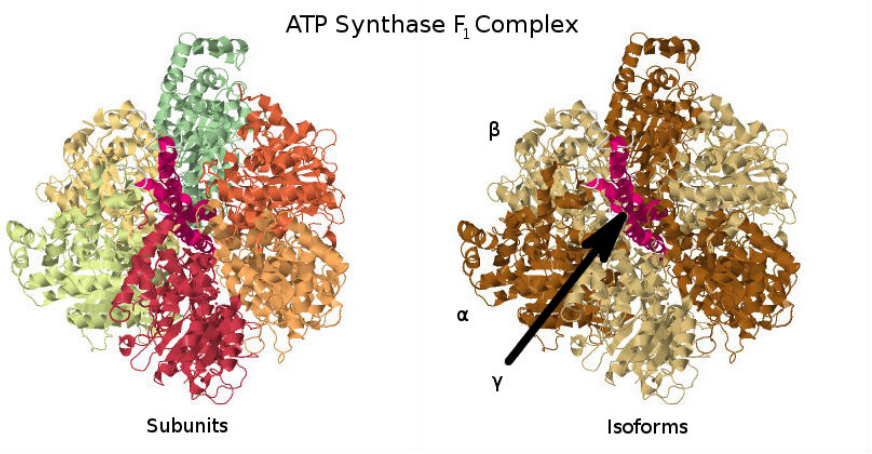


Figure C.6: Illustration of the F_1 part of the ATP Synthase complex (PDB ID 1E79; Gibbons et al. 4, Bernstein et al. 5, Gezelter et al. 6). This illustration demonstrates both how an enzyme complex may be constituted by multiple subunits (left), and how some of those subunits may be products of the same gene and have differing stoichiometries within the complex (right).

Assumption 3. The modeling of enzyme complex abundance can be tackled by using nested sets of subcomplexes; each enzyme complex consists of multiple subcomplexes, unless it is only a single protein or family of protein isozymes. These subcomplexes are required for the enzyme complex to function (AND relationships), and can be thought of as the division of the complex in to distinct units that each have some necessary function for the complex, with the exception that we do not keep track of the multiplicity of subcomplexes within a complex since this information is, in the current state of affairs, not always known. However, there may be alternative versions of each functional set (given by OR relationships). Eventually, this nested embedding terminates with a single protein or set of peptide isoforms (e.g. isozymes). In the case of ATP Synthase, one of its functional sets is represented by the F_1 subcomplex. The F_1 subcomplex itself can be viewed as having two immediate subcomplexes: the single γ (axle) subunit and three identical subcomplexes each made of an α and β subunit. Each $\alpha\beta$ pair works together to bind ADP and catalyze the reaction [191]. The $\alpha\beta$ subcomplex itself then has two subcomplexes composed of just an α subunit on the one hand and the β subunit

on the other. It is obvious that one of these base-level functional subcomplexes (in this example, either γ or $\alpha\beta$) will be in most limited supply, and that it will best represent the overall enzyme complex abundance (discounting the issues of multiplicity for $\alpha\beta$, see Assumption 4).

The hierarchical structure just described, when written out in Boolean, will give a rule in CNF (conjunctive normal form), or more specifically (owing to the lack of negations), clausal normal form, where a clause is a disjunction of literals (genes). This is because all relations are ANDs (conjunctions), except possibly at the inner-most subcomplexes that have alternative isoforms, which are expressed as ORs (disjunctions). Since GPR rules alone only specify the requirements for enzyme complex formation, we will see that not all forms of Boolean rules are equally useful in evaluating the enzyme complex abundance, but we have established the assumptions in Table C.1 and an alternative and logically equivalent rule [160] under which we can estimate enzyme complex copy number.

There is no guarantee that a GPR rule has been written down with this hierarchical structure in mind, though it is likely the case much of the time as it is a natural way to model complexes. However, any GPR rule can be interpreted in the context of this hierarchical view due to the existence of a logically equivalent CNF rule for any non-CNF rule, and it is obvious that logical equivalence is all that is required to check for enzyme complex formation when exact isoform stoichiometry is unknown. As an example, we consider another common formulation for GPR rules, and a way to think about enzyme structure—disjunctive normal form (DNF). A DNF rule is a disjunctive list of conjunctions of peptide isoforms, where each conjunction is some variation of the enzyme complex due to substituting in different isoforms for some of the required subunits. A rule with a more complicated structure and compatible isoforms across sub-

Table C.1: A list of assumptions about how Gene-Protein-Reaction rules can describe enzyme complex stoichiometry.

Table C.1. Assumptions in GPR-based Enzyme Complex Formation

1. Expression values are highly correlated with the copy numbers of their corresponding peptide isoforms.
2. Protein isoforms contributing to isozymes are considered part of the same enzyme complex.
3. Any enzyme complex can be described as a hierarchical subset of (possibly redundant) subcomplexes; redundant subcomplexes, as elaborated in (4), are not currently modeled.
4. Assume one copy of peptide per complex; exact isoform stoichiometry is not considered.
5. With the exception of complexes having identical rules (i.e. the same complex listed for different reactions), each copy of a peptide is available for all complexes in the model.
6. There is only one active site per enzyme complex.
7. We assume that different pathways have similar flux sensitivities with respect to their enzyme abundances.
8. If a particular subcomplex can be catalyzed by A and it can also be catalyzed by A and B (e.g. B acts as a regulatory unit, as in holoenzymes), this just simplifies to A once expression values are substituted in. Similarly, allosteric regulation is not modeled. Relatedly, there are no NOT operations in GPR rules (just ANDs and ORs).
9. Enzyme complexes form without the assistance of protein chaperones and formation is not coupled to other reactions.
10. Post-translational modifications do not affect complex formation.
11. Rate of formation and degradation of complexes doesn't play a role, since we assume steady-state.

complexes may be written more succinctly in CNF, whereas a rule with only very few alternatives derived from isoform variants may be represented clearly with DNF. In rare cases, it is possible that a GPR rule is written in neither DNF or CNF, perhaps because neither of these two alternatives above are strictly the case, and some other rule is more succinct.

Assumptions 4, 5 and 6. One active site per enzyme complex implies a single complex can only catalyze one reaction at a time. Multimeric complexes with one active site per identical subunit would be considered as one enzyme complex per subunit in this model. Note that it is possible for an enzyme complex to catalyze different reactions. In fact, some transporter complexes can transfer many different metabolites across a lipid bilayer—up to 294 distinct reactions in the reversible model for solute carrier family 7 (Gene ID 9057). Another example is the ligation or hydrolysis of nucleotide, fatty acid, or peptide chains, where chains of different length may all be substrates or products of the same enzyme complex. While we do not explicitly consider these in in the minimum disjunction algorithm, these redundancies are taken into account subsequently in Algorithm 1.

What is currently not considered in our process is that some peptide isoforms may find use in completely different complexes, and in some cases, individual peptides may have multiple active sites; in the first case, we assume an unrealistic case of superposition where the isoform can simultaneously function in more than one complex. The primary reason we have not tackled this problem is because exact subunit stoichiometry of most enzyme complexes is not accurately known, but an increasing abundance of data on BRENDA [192] gives some hope to this problem. A recent *E. coli* metabolic model incorporating the metabolism of all known gene products [187] also includes putative enzyme complex stoichiometry in GPR rules. For the second point, there are a

few enzymes where a single polypeptide may have multiple active sites (e.g. fatty acid synthase), and this is not currently taken into account in our model.

Assumption 8. We do not make any special assumptions requiring symmetry of an isoform within a complex. For instance, the example in assumption 8 shows how you might have one subcomponent composed of a single isoform, and another subcomponent composed of that gene in addition to another isoform. In this case, it is simply reduced to being the first gene only that is required, since clearly the second is strictly optional. That isn't to say that the second gene may not have some metabolic effect, such as (potentially) aiding in structural ability or altering the catalytic rate, but it should have no bearing on the formation of a functional catalytic complex. Holoenzymes—enzymes with metabolic cofactors or protein subunits that have a regulatory function for the complex—would likely be the only situation where this type of rule might need to be considered in more detail. But in the absence of detailed kinetic information, this consideration (much like allosteric regulation) is not useful.

No additional algorithmic considerations are needed, as this is a by-product of the conversion to CNF. For instance, take the following example where the second conjunction has the redundant gene g_3:

$$(g_1 \wedge g_2) \vee (g_1 \wedge g_2 \wedge g_3)$$

Distributing during the process of conversion to CNF results in:

$$g_1 \wedge (g_1 \vee g_2) \wedge (g_1 \vee g_3) \wedge (g_2 \vee g_1) \wedge g_2 \wedge (g_2 \vee g_3)$$

Because every disjunction with more than one literal is in conjunction with another

disjunction with only one of its literals, the disjunction with fewer literals will be the minimum of the two once evaluated. This applies to both of the singleton disjunctions g_1 and g_2, so all other disjunctions will effectively be ignored (it is up to the implementer whether the redundant sub-expressions are removed before evaluation):

$$g_1 \wedge \cancel{(g_1 \vee g_2)} \wedge \cancel{(g_1 \vee g_3)} \wedge \cancel{(g_2 \vee g_1)} \wedge g_2 \wedge \cancel{(g_2 \vee g_3)} = (g_1 \wedge g_2)$$

Assumption 7. Another important biochemical assumption is that reactions should operate in a regime where they are sensitive to changes in the overall enzyme level in the pathways that they belong in [166, 167]. This is perhaps the most important issue to be explored further for methods like this, since if it is not true, some other adjustment factor would be needed to make the method realistic. For instance, if all reactions in a pathway are operating far below V_{max}, but it is not the case in another pathway, the current method does not have information on this, and will try to put more flux through the first pathway than should be the case.

Assumptions 9, 10 and 11. Due to the quickness, stability, and energetic favorability of enzyme complex formation, the absence of chaperones or coupled metabolic reactions required for complex formation may be reasonable assumptions, but further research is warranted [68]. Additionally, as in metabolism, we assume a steady state for complex formation, so that rate laws regarding complex formation aren't needed. However, further research may be warranted to investigate the use of a penalty for complex levels based on mass action and protein-docking information. Requisite to this would be addressing assumption 4. It would be surprising (but not impossible) if such a penalty were very large due to the cost this would imply for many of the large and important enzyme complexes present in all organisms [193]. A more serious consideration may be that information on post-translational modification is not currently considered. Post-

translational modification is highly context-specific and the relevant data is not as cheap to get as expression data, so it may be some time before it can be integrated into the modeling framework.

Table C.2: Running times (in seconds, ± standard deviation) for FALCON using various algorithms implemented in the Gurobi package. For yeast models, 1,000 replicates were performed, and for the human model, 100 replicates were performed.

Model	*Primal-Simplex*	*Dual-Simplex*	*Barrier*
Yeast 5.21 (2,061 reactions)	7.841 ± 1.697	7.611 ± 1.267	10.859 ± 2.788
Yeast 7.0 (3,498 reactions)	51.863 ± 22.731	65.317 ± 12.771	242.137 ± 57.129
Human 2.03 (7,440 reactions)	159.077 ± 24.903	152.297 ± 39.783	366.166 ± 92.321

C.3 Benchmarking of solvers

We have exclusively used the Gurobi solver [194] for this work, which is a highly competitive solver that employs by default a parallel strategy to solving problems: a different algorithm is run simultaneously, and as soon as one algorithm finished the others terminate. Of course, if there is a clear choice of algorithm for a particular problem class, this should be used in production settings to avoid wasted CPU time and memory. In order to address this, we benchmarked the three non-parallel solver methods in Gurobi (since parallel solvers simply use multiple methods simultaneously). The exception to this rule is the Barrier method, which can use multiple threads, but in practice for our models appears to use no more than about 6 full CPU cores simultaneously for our models. Our results for Yeast 5 and Yeast 7 with minimal directionality constraints [1, 162, 165] and Human Recon 2 [63] are shown in Table C.2).

We found that in Yeast 7 with the primal-simplex solver, there is a chance the solver will fail to find a feasible solution. We verified that this is a numeric issue in Gurobi and can be fixed by setting the Gurobi parameter `MarkowitzTol` to a larger value (which decreases time-efficiency but limits the numerical error in the simplex algorithm). In practice, failure for the algorithm to converge at an advanced iteration is rare and is not always a major problem (since the previous flux estimate by the advanced iteration should already be quite good), but it is certainly undesirable; a warning message will be

Table C.3: Running time per FALCON iteration (in seconds, ± standard deviation) using various algorithms implemented in the Gurobi package. For yeast models, 1,000 replicates were performed, and for the human model, 100 replicates were performed.

Model	*Primal-Simplex*	*Dual-Simplex*	*Barrier*
Yeast 5.21 (2,061 reactions)	0.721 ± 0.023	0.652 ± 0.040	1.100 ± 0.112
Yeast 7.0 (3,498 reactions)	2.725 ± 0.298	2.469 ± 0.289	11.309 ± 1.589
Human 2.03 (7,440 reactions)	6.422 ± 0.484	5.233 ± 0.661	15.782 ± 3.209

printed by `falcon` if this occurs, at which point parameter settings can be investigated. In the future, we plan to improve `falcon` so that parameters will be adjusted as needed during progression of the algorithm after finding a good test suite of models and data. For now, we use the dual-simplex solver, for which we have always had good results.

Because the number of iterations depends non-trivially on the model and the expression data, it may be more helpful to look at the average time per iteration in the above examples (Table C.3).

Given the above rare trouble with primal simplex solver the universal best performance enjoyed by the dual-simplex method (Tables C.2 and C.3), we would advise the dual-simplex algorithms, all else being equal. The dual-simplex method is also recommended for memory-efficiency by Gurobi documentation, but we did not observe any differences in memory for different solver methods.

All timing analyses were performed on a system with four 8-core AMD Opteron™ 6136 processors operating at 2.4 GHz. Figure C.2, Table 4.1, and Tables C.2 and C.3 used a single unperturbed expression file per species (*S. cerevisiae* and *H. sapiens*; see `timingAnalysis.m` for details). Values were averaged across 32 replicates. Note that the iMAT method is formulated as a mixed integer program [30], and was able to use additional parallelization of the solver [194] whereas other methods only used a single core (our system had 32 cores and iMAT with Gurobi would use all of them). Tables C.2

and C.3 used multivariate log-normal noise multiplied by the original expression vector to introduce more variance in the calculations; the human models were tested with 100 replicates and the yeast models with 500 replicates.

C.4 Generation of figures and tables

All non-trivial figures can be generated using MATLAB scripts found in the `analysis/figures` subdirectory of the FALCON installation. In particular, figures should be generated through the master script `makeMethodFigures.m` by calling `makeMethodFigures(figName)` where `figName` has a name corresponding to the desired figure. In some cases, some MATLAB `.mat` files will need to be generated by other scripts first; see the plotting scripts or the subsections below for details. An example is to make the scatter plots showing the difference between running falcon with enzyme abundances determined by direct evaluation or the minimum disjunction algorithm; all three scatter plots are generated with the command `makeMethodFigures('fluxCmpScatter')`. Note that, as written, this requires a graphical MATLAB session.

Comparison of the effects of the employed enzyme complexation methods were evaluated using `compareEnzymeExpression.m` and `compareFluxByRGroup.m`. Comparison of reaction groups was performed in `compareFluxByRGroup.m`.

C.4.1 Timing analyses

All timing analyses were performed on a system with four 8-core AMD Opteron™ 6136 processors operating at 2.4 GHz. Figure C.2 and Table 4.1 used unperturbed expression

data; see `yeastResults.m` for details). Values for the FALCON method were averaged across 32 replicates, while values for the Lee et al. 1 method were averaged across 8 replicates. Human timing analyses were performed using `methodTimer.m` with 8 replicates.

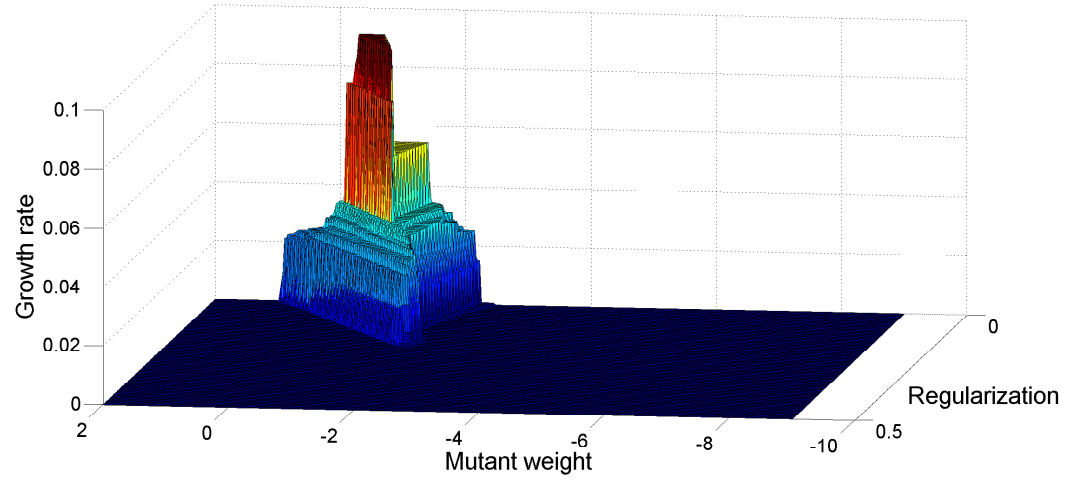
C.4.2 Data sources

Enzyme complexation comparisons were performed on proteomics data from Gholami et al. 164 (Human; 786-O cell line) and Picotti et al. 163 (yeast; BY strain), and on RNA-Seq data from Lee et al. 1 (yeast; 75% max μ condition).

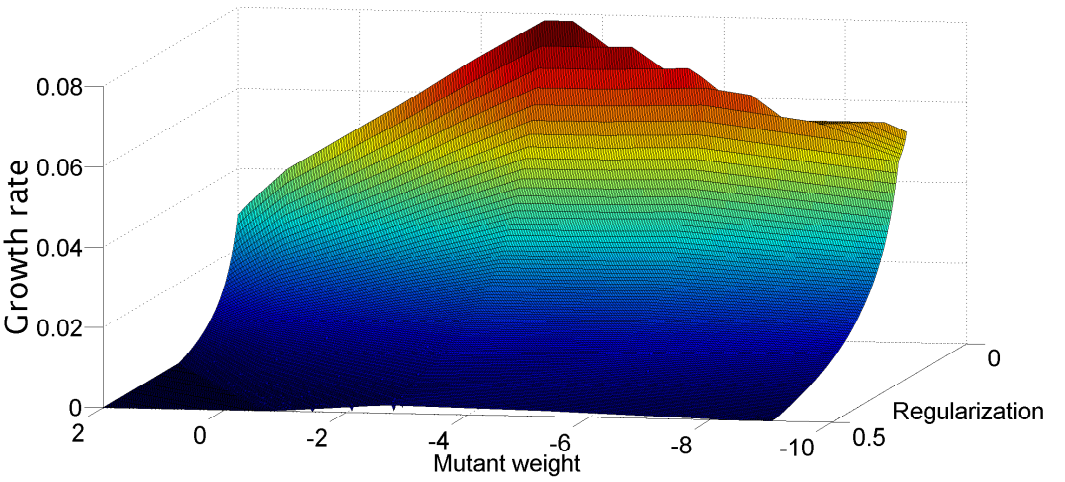
APPENDIX D

SIMULATION OF BENEFICIAL MUTATIONS

D.1 Supporting figures



(a) Linear MoMA



(b) Quadratic MoMA

Figure D.1: The same reaction is used in both figures, and in both instances, a slightly negative weight on the reaction appears to be most beneficial (compare to 0, which represents the wild-type). Weight on a (linear or quadratic, respectively) regularization objective component is shown on the y-axis, which is often a helpful constant both biologically and for removing invalid flux cycles [7, 8].

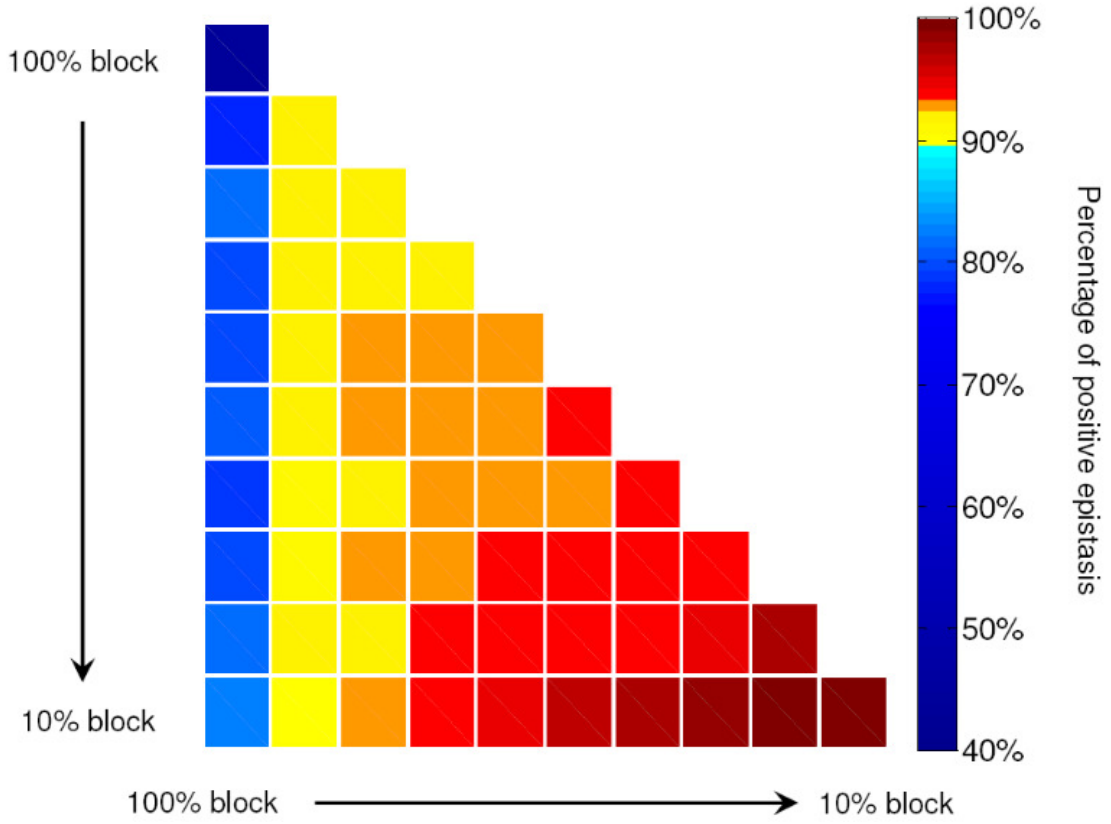
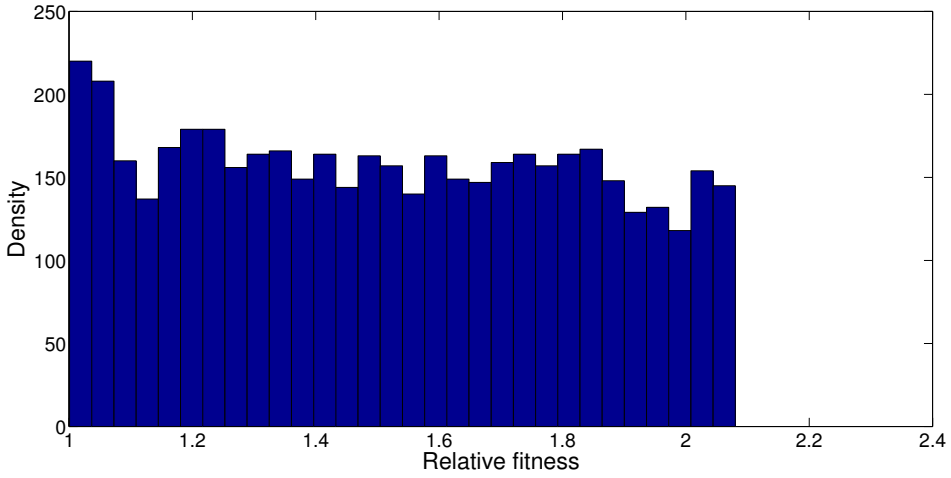
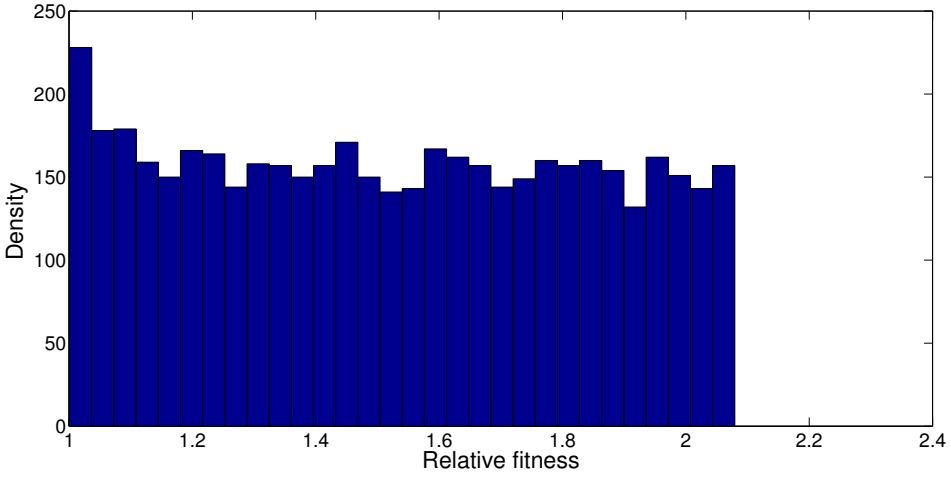


Figure D.2: Flux restriction increases the percentage of negative epistatic interactions. Data taken from Xu et al. [9].



(a) 50% Gaussian sampling



(b) Uniform sampling

Figure D.3: Example distributions of beneficial mutations for the yeast YPE example when sampling 50% of flux mutations within the larger FVA bound using a truncated normal distribution (D.3a) or when using uniform sampling between the FVA bounds (D.3b).

D.2 Supporting information

D.2.1 Evolutionary path analysis

The repository housing the project is currently located at: `https://github.com/bbarker/COBRAscripts`. The C code which is used for the analysis may be found in the `MyProjects/AdaptiveMuts/TreeTraversal` subdirectory.

The power-set of mutations can be ordered in the natural binary order. For instance, four mutations can be ordered from 0000 (wild-type) to 1111 (all mutations present) , where a '0' denotes the absence of the given mutation and a '1' denotes its presence. An example of the four mutation case can be given in one line of a text file as follows:

```
0.013743, 0.024794, 0.020672, 0.023515, 0.019147, 0.02291, 0, 0, 0.017884,
0.024196, 0.016066, 0.023515, 0.015166, 0.02291, 0, 0
```

The first fitness in the list, which corresponds to the wild-type, is the divisor in the following output from `printPaths`[1], showing all 4! paths that may arise from the above mutants.

```
*** Replicate 0 ***
0       4       3       2       1
1.000000        1.301317        1.103544        0.000000        0.000000
0       4       3       1       2
1.000000        1.301317        1.103544        1.667030        0.000000
0       4       2       3       1
1.000000        1.301317        1.169032        0.000000        0.000000
```

[1] `printpaths` has the same usage as `trapFind`; see page 150

```
0       4       2       1       3
1.000000        1.301317        1.169032        1.711053        0.000000
0       4       1       3       2
1.000000        1.301317        1.760605        1.667030        0.000000
0       4       1       2       3
1.000000        1.301317        1.760605        1.711053        0.000000
0       3       4       2       1
1.000000        1.393218        1.103544        0.000000        0.000000
0       3       4       1       2
1.000000        1.393218        1.103544        1.667030        0.000000
0       3       2       4       1
1.000000        1.393218        0.000000        0.000000        0.000000
0       3       2       1       4
1.000000        1.393218        0.000000        0.000000        0.000000
0       3       1       4       2
1.000000        1.393218        1.667030        1.667030        0.000000
0       3       1       2       4
1.000000        1.393218        1.667030        0.000000        0.000000
0       2       4       3       1
1.000000        1.504184        1.169032        0.000000        0.000000
0       2       4       1       3
1.000000        1.504184        1.169032        1.711053        0.000000
0       2       3       4       1
1.000000        1.504184        0.000000        0.000000        0.000000
0       2       3       1       4
1.000000        1.504184        0.000000        0.000000        0.000000
```

```
0       2       1       4       3
1.000000        1.504184        1.711053        1.711053        0.000000
0       2       1       3       4
1.000000        1.504184        1.711053        0.000000        0.000000
0       1       4       3       2
1.000000        1.804118        1.760605        1.667030        0.000000
0       1       4       2       3
1.000000        1.804118        1.760605        1.711053        0.000000
0       1       3       4       2
1.000000        1.804118        1.667030        1.667030        0.000000
0       1       3       2       4
1.000000        1.804118        1.667030        0.000000        0.000000
0       1       2       4       3
1.000000        1.804118        1.711053        1.711053        0.000000
0       1       2       3       4
1.000000        1.804118        1.711053        0.000000        0.000000
```

'Replicate 0' merely denotes that this output was due to the first example power set of mutants in a file, since `randomBeneficialPsets.m` or `randomAdaptivePsets.m` can be used to generate multiple sets of mutants. The primary difference between the two scripts is that the former uses hard constraint changes as mutations, whereas the latter uses weights (which is the version used in the present study). In an experimental setting, it is much more likely that we would only have one replicate, which is exactly why experimental settings are unlikely to be useful for showing trends in adaptive evolution. Each line of fitnesses is preceded by a corresponding line which lists which mutation was added at that point in the evolutionary path (0 to n where n is the number of mutations, so $n = 4$ in this example). It may be desirable to map these mutations back to the reactions

Table D.1: Example output of `trapFind` applied to experimental datasets. The input files may be found in this project's `TreeTraversal/Experiments` subdirectory.

study	mutations	reached optima	trap: fitness decrease	trap: local maxima
Weinreich et al. [174]	5	96	18	0
Khan et al. [178]	5	108	7	0
Chou et al. [173]	4	24	0	0

they are associated with in the model. This output from the above scripts is stored in files with the suffix `_rxnlist.csv`.

The `trapFind` program can be used to find how many paths have traps (local fitness optima or fitness decreases) that prevent adaptation from reaching the global fitness maximum among all combinations of mutations in the set.

This program takes as its first argument a file containing comma-delimited lists of mutant fitnesses described above, followed by the number of individual mutations as the second argument, and the number of replicates (lines in the file) as the third argument:

```
trapFind [number of mutants] [number of replicates]
```

The output is contained in `[original file prefix]_trapAnalysis.csv`. The contents of the file are comma-separated lists (one per mutation set) that have the following information: number of mutations, number of paths that reach the optimum, number paths terminated due to a decrease in fitness, and the number of paths terminated due to local maxima. As a further example, the output for three experimental datasets has been listed (Table D.1).

Note that some paths that become trapped may have themselves branched into other paths, so the tally for the last three columns does not necessarily reach $n!$. If we consider that neutral mutations are unlikely to become fixed in a population [174], only 18 of 120 paths are evolutionarily inaccessible, instead of the 96 encountered in this analysis.

D.2.2 Pairwise adaptive mutations

The function we used to generate the pairwise epistasis data is `randomEpistasisSampler` found in `MyProjects/AdaptiveMuts`. This function relies on another function in the same directory for generating a pool of single mutants, `randomSingleBeneMuts`.

BIBLIOGRAPHY

[1] Dave Lee, Kieran Smallbone, Warwick B Dunn, Ettore Murabito, Catherine L Winder, Douglas B Kell, Pedro Mendes, and Neil Swainston. Improving metabolic flux predictions using absolute gene expression data. *BMC Syst. Biol.*, 6(1):73, January 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-73. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3477026&tool=pmcentrez&rendertype=abstract.

[2] Alexander A Shestov, Brandon Barker, Zhenglong Gu, and Jason W Locasale. Computational approaches for understanding energy metabolism. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 5(6):733–750, July 2013. ISSN 1939-005X. doi: 10.1002/wsbm.1238. URL http://dx.doi.org/10.1002/wsbm.1238.

[3] Mohit Jain, Roland Nilsson, Sonia Sharma, Nikhil Madhusudhan, Toshimori Kitami, Amanda L Souza, Ran Kafri, Marc W Kirschner, Clary B Clish, and Vamsi K Mootha. Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation. *Science (80-.).*, 336(6084):1040–1044, May 2012. doi: 10.1126/science.1218595. URL http://www.sciencemag.org/content/336/6084/1040.abstract.

[4] Clyde Gibbons, Martin G Montgomery, Andrew G W Leslie, and John E Walker. The structure of the central stalk in bovine F1-ATPase at 2.4 A resolution. *Nat Struct Mol Biol*, 7(11):1055–1061, November 2000. ISSN 1072-8368. URL http://dx.doi.org/10.1038/80981.

[5] Frances C Bernstein, Thomas F Koetzle, Graheme J B Williams, Edgar F Meyer Jr., Michael D Brice, John R Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.*, 185(2):584–591, January 1978. ISSN 0003-9861. doi: http://dx.doi.org/10.1016/0003-9861(78)90204-7. URL http://www.sciencedirect.com/science/article/pii/0003986178902047.

[6] Dan Gezelter, Bradley A Smith, and Egon Willighagen. Jmol: an open-source Java viewer for chemical structures in 3D, 2013. URL http://www.jmol.org/.

[7] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer. Multidimensional Optimality of Microbial Metabolism. *Science (80-.).*, 336(6081):601–604, May 2012. ISSN 0036-8075. doi: 10.1126/science.1216882. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1216882.

[8] Kieran Smallbone and Evangelos Simeonidis. Flux balance analysis: a geometric perspective. *J. Theor. Biol.*, 258(2):311–5, 2009. ISSN 1095-8541. doi: 10.1016/j.jtbi.2009.01.027. URL http://www.ncbi.nlm.nih.gov/pubmed/19490860.

[9] Lin Xu, Brandon Barker, and Zhenglong Gu. Dynamic epistasis for different alleles of the same gene. *Proc. Natl. Acad. Sci. U. S. A.*, June 2012. ISSN 1091-6490. doi: 10.1073/pnas.1121507109. URL http://www.ncbi.nlm.nih.gov/pubmed/22689976.

[10] Howard M Shapiro. *Studies in the Structure of the Bacterial Economy: An Input-Output Model of Escherichia Coli*. PhD thesis, Harvard University, 1961.

[11] Howard M Shapiro. Input-Output Models of Biological Systems: Formulation and Applicability. *Comput. Biomed. Res.*, 2:430–445, 1969.

[12] C Van De Panne and F Rahnamat. The First Algorithm for Linear Programming : An Analysis of Kantorovich ' s Method. *Econ. Plan.*, 19 (2):76–91, 1985.

[13] M.R. Watson. Metabolic maps for the Apple II. *Biochem. Soc. Trans.*, 12:1093–1094, 1984.

[14] J M Savinell and Bernhard Ø Palsson. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J. Theor. Biol.*, 154(4):421–54, March 1992. ISSN 0022-5193. URL http://www.ncbi.nlm.nih.gov/pubmed/1593896.

[15] Joanne M Savinell, Bernhard Ø Palsson, and Ann Arbor. Network Analysis of Intermediary Metabolism using Linear Optimization . II . Interpretation of Hybridoma Cell Metabolism The uses of linear optimization theory to calculate and interpret fluxes in metabolic. *Cell Metab.*, pages 455–473, 1992.

[16] J S Edwards and Bernhard Ø Palsson. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U. S. A.*, 97(10):5528–33, May 2000. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=25862&tool=pmcentrez&rendertype=abstract.

[17] Iman Famili, Jochen Forster, Jens Nielsen, and Bernhard Ø Palsson. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. U. S. A.*, 100(23):13134–9, November 2003. ISSN 0027-8424. doi: 10.1073/pnas.2235812100. URL http://www.ncbi.nlm.nih.gov/pubmed/14578455.

[18] Stephen S Fong and Bernhard Ø Palsson. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat. Genet.*, 36(10):1056–8, October 2004. ISSN 1061-4036. doi: 10.1038/ng1432. URL http://www.ncbi.nlm.nih.gov/pubmed/15448692.

[19] Ronan M T Fleming, C M Maes, M a Saunders, Y Ye, and Bernhard Ø Palsson. A variational principle for computing nonequilibrium fluxes and potentials in genome-scale biochemical networks. *J. Theor. Biol.*, 292:71–7, January 2012. ISSN 1095-8541. doi: 10.1016/j.jtbi.2011.09.029. URL http://www.ncbi.nlm.nih.gov/pubmed/21983269.

[20] A Varma and Bernhard Ø Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.*, 60(10):3724–31, October 1994. ISSN 0099-2240. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=201879&tool=pmcentrez&rendertype=abstract.

[21] M M Zavlanos and A A Julius. Robust flux balance analysis of metabolic networks. In *Am. Control Conf. (ACC), 2011*, pages 2915–2920, June 2011. doi: 10.1109/ACC.2011.5991248.

[22] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nat. Biotechnol.*, 28(3):245–8, March 2010. ISSN 1546-1696. doi: 10.1038/nbt.1614. URL http://www.ncbi.nlm.nih.gov/pubmed/20212490.

[23] Daniel Segrè, Dennis Vitkup, and George M Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.*, 99(23):15112–7, November 2002. ISSN 0027-8424. doi: 10.1073/pnas.232349399. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=137552&tool=pmcentrez&rendertype=abstract.

[24] Tomer Shlomi, Omer Berkman, and Eytan Ruppin. Regulatory on/off minimization of metabolic flux. *Proc. Natl. Acad. Sci.*, 102(21):7695–7700, 2005.

[25] Jan Schellenberger, Nathan E Lewis, and Bernhard Ø Palsson. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophys. J.*, 100(3):544–53, February 2011. ISSN 1542-0086. doi: 10.1016/j.bpj.2010.12.3707. URL http://www.ncbi.nlm.nih.gov/pubmed/21281568.

[26] D Beard, S Liang, and H Qian. Energy Balance for Analysis of Complex Metabolic Networks. *Biophys. J.*, 83(1):79–86, July 2002. ISSN 00063495. doi: 10.1016/S0006-3495(02)75150-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0006349502751503.

[27] Christopher S Henry, Linda J Broadbelt, and Vassily Hatzimanikatis. Thermodynamics-based metabolic flux analysis. *Biophys. J.*, 92(5):1792–805, March 2007. ISSN 1542-0086. doi: 10.1529/biophysj.106.093138. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1796839&tool=pmcentrez&rendertype=abstract.

[28] Yuliang Wang, James a Eddy, and Nathan D Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.*, 6:153, January 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-153. URL http://www.ncbi.nlm.nih.gov/pubmed/23234303.

[29] Scott A Becker and Bernhard Ø Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.*, 4(5):1–10, May 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000082. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2366062&tool=pmcentrez&rendertype=abstract.

[30] Tomer Shlomi, Moran N Cabili, Markus J Herrgård, Bernhard Ø Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, 26(9):1003–10, September 2008. ISSN 1546-1696. doi: 10.1038/nbt.1487. URL http://www.ncbi.nlm.nih.gov/pubmed/18711341.

[31] Sriram Chandrasekaran and Nathan D Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.*, 107(41):17845–50, October 2010. ISSN 1091-6490. doi: 10.1073/pnas.1005139107. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2955152&tool=pmcentrez&rendertype=abstract.

[32] Marcin Imielinski and Calin Belta. Deep epistasis in human metabolism. *Chaos*, 20(2):026104, June 2010. ISSN 1089-7682. doi: 10.1063/1.3456056. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2909311&tool=pmcentrez&rendertype=abstract.

[33] Marcin Imielinski and Calin Belta. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Syst. Biol.*, 2:40, January 2008. ISSN 1752-0509. doi: 10.1186/1752-0509-2-40. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2390508&tool=pmcentrez&rendertype=abstract.

[34] Xionglei He, Wenfeng Qian, Zhi Wang, Ying Li, and Jianzhi Zhang. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat. Genet.*, 42(3):272–6, 2010. ISSN 1546-1718. doi: 10.1038/ng.524. URL http://www.ncbi.nlm.nih.gov/pubmed/20101242.

[35] Christopher S Henry, Matthew D Jankowski, Linda J Broadbelt, and Vassily Hatzimanikatis. Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophys. J.*, 90(4):1453–61, February 2006. ISSN 0006-3495. doi: 10.1529/biophysj.105.071720. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1367295&tool=pmcentrez&rendertype=abstract.

[36] Anne Kümmel, Sven Panke, and Matthias Heinemann. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.*, 2:2006.0034, January 2006. ISSN 1744-4292. doi: 10.1038/msb4100074. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1681506&tool=pmcentrez&rendertype=abstract.

[37] Arne C Müller and Alexander Bockmayr. Fast thermodynamically constrained flux variability analysis. *Bioinformatics*, pages 1–7, February 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt059. URL http://www.ncbi.nlm.nih.gov/pubmed/23390138.

[38] H S Haraldsdóttir, Ines Thiele, and Ronan M T Fleming. Quantitative assignment of reaction directionality in a multicompartmental human metabolic reconstruction. *Biophys. J.*, 102(8):1703–11, April 2012. ISSN 1542-0086. doi: 10.1016/j.bpj.2012.02.032. URL http://www.ncbi.nlm.nih.gov/pubmed/22768925.

[39] Daniele De Martino, Matteo Figliuzzi, Andrea De Martino, and Enzo Marinari. A scalable algorithm to explore the Gibbs energy landscape of genome-scale metabolic networks. *PLoS Comput. Biol.*, 8(6):e1002562, January 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002562. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3380848&tool=pmcentrez&rendertype=abstract.

[40] Prabhasa Ravikirthi, Patrick F Suthers, and Costas D Maranas. Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnol. Bioeng.*, 108(6):1372–82, June 2011. ISSN 1097-0290. doi: 10.1002/bit.23070. URL http://www.ncbi.nlm.nih.gov/pubmed/21328316.

[41] Patrick Warren and Janette Jones. Duality, Thermodynamics, and the Linear Programming Problem in Constraint-Based Models of Metabolism. *Phys. Rev. Lett.*, 99(10):108101, September 2007. ISSN 0031-9007. doi: 10.1103/PhysRevLett.99.108101. URL http://link.aps.org/doi/10.1103/PhysRevLett.99.108101.

[42] Nathan E Lewis, Gunnar Schramm, Aarash Bordbar, Jan Schellenberger, Michael P Andersen, Jeffrey K Cheng, Nilam Patel, Alex Yee, Randall a Lewis, Roland Eils, Rainer König, and Bernhard Ø Palsson. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.*, 28(12):1279–85, December 2010. ISSN 1546-1696. doi: 10.1038/nbt.1711. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3140076&tool=pmcentrez&rendertype=abstract.

[43] Neema Jamshidi and Bernhard Ø Palsson. Systems biology of the human red blood cell. *Blood Cells. Mol. Dis.*, 36(2):239–47, 2006. ISSN 1079-9796. doi: 10.1016/j.bcmd.2006.01.006. URL http://www.ncbi.nlm.nih.gov/pubmed/16533612.

[44] Yuekai Sun, Ronan M T Fleming, Ines Thiele, and Michael A Saunders. Robust flux balance analysis of multiscale biochemical reaction networks. 2012.

[45] Aarash Bordbar, Nathan E Lewis, Jan Schellenberger, Bernhard Ø Palsson, and Neema Jamshidi. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol. Syst. Biol.*, 6(422):422, October 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.68. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2990636&tool=pmcentrez&rendertype=abstract.

[46] Niels Klitgord and Daniel Segrè. Ecosystems biology of microbial metabolism. *Curr. Opin. Biotechnol.*, pages 541–546, May 2011. ISSN 1879-0429. doi: 10.1016/j.copbio.2011.04.018. URL http://www.ncbi.nlm.nih.gov/pubmed/21592777.

[47] Samuel M D Seaver, Christopher S Henry, and Andrew D Hanson. Frontiers in metabolic reconstruction and modeling of plant genomes. *J. Exp. Bot.*, 63(6):2247–58, March 2012. ISSN 1460-2431. doi: 10.1093/jxb/err371. URL http://www.ncbi.nlm.nih.gov/pubmed/22238452.

[48] Ramy K Aziz, Scott Devoid, Terrence Disz, Robert a Edwards, Christopher S Henry, Gary J Olsen, Robert Olson, Ross Overbeek, Bruce Parrello, Gordon D Pusch, Rick L Stevens, Veronika Vonstein, and Fangfang Xia. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS One*, 7(10):e48053, January 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0048053. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3480482&tool=pmcentrez&rendertype=abstract.

[49] Christopher S Henry, Ross Overbeek, Fangfang Xia, Aaron a Best, Elizabeth Glass, Jack Gilbert, Peter Larsen, Rob Edwards, Terry Disz, Folker Meyer, Veronika Vonstein, Matthew Dejongh, Daniela Bartels, Narayan Desai, Mark D'Souza, Scott Devoid, Kevin P Keegan, Robert Olson, Andreas Wilke, Jared Wilkening, and Rick L Stevens. Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochim. Biophys. Acta*, 1810(10):967–77, October 2011. ISSN 0006-3002. doi: 10.1016/j.bbagen.2011.03.010. URL http://www.ncbi.nlm.nih.gov/pubmed/21421023.

[50] José P Faria, Ross Overbeek, Fangfang Xia, Miguel Rocha, Isabel Rocha, and Christopher S Henry. Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. *Brief. Bioinform.*, pages 1–20, February 2013. ISSN 1477-4054. doi: 10.1093/bib/bbs071. URL http://www.ncbi.nlm.nih.gov/pubmed/23422247.

[51] Jason W Locasale. Metabolic rewiring drives resistance to targeted cancer therapy. *Mol. Syst. Biol.*, 8(597):597, January 2012. ISSN 1744-4292. doi: 10.1038/msb.2012.30. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3421445&tool=

pmcentrez&rendertype=abstract.

[52] Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.*, 7(501):501, January 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.35. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159974&tool=pmcentrez&rendertype=abstract.

[53] Markus W Covert, C H Schilling, and Bernhard Ø Palsson. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.*, 213(1):73–88, 2001. ISSN 0022-5193. doi: 10.1006/jtbi.2001.2405. URL http://www.ncbi.nlm.nih.gov/pubmed/11708855.

[54] Markus J Herrgård, Baek-Seok Lee, Vasiliy Portnoy, and Bernhard Ø Palsson. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. *Genome Res.*, 16(5):627–35, 2006. ISSN 1088-9051. doi: 10.1101/gr.4083206. URL http://www.ncbi.nlm.nih.gov/pubmed/16606697.

[55] Janis Dingel and Olgica Milenkovic. List-decoding methods for inferring polynomials in finite dynamical gene network models. *Bioinformatics*, 25(13):1686–93, July 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp281. URL http://www.ncbi.nlm.nih.gov/pubmed/19401400.

[56] Elena Dimitrova, Franziska Hinkelmann, Abdul S, Reinhard Laubenbacher, Brandilyn Stigler, and Michael Stillman. Parameter estimation for Boolean models of biological networks. *Theor. Comput. Sci.*, 412:2816–2826, 2011.

[57] B Stigler, a Jarrah, M Stillman, and R Laubenbacher. Reverse engineering of dynamic networks. *Ann. N. Y. Acad. Sci.*, 1115(0477):168–77, December 2007. ISSN 0077-8923. doi: 10.1196/annals.1407.012.

[58] Abdul Salam Jarrah, Reinhard Laubenbacher, Brandilyn Stigler, and Michael Stillman. Reverse-engineering of polynomial dynamical systems. *Adv. Appl. Math.*, 39:477–489, 2007.

[59] Osbaldo Resendis-Antonio, Alberto Checa, and Sergio Encarnación. Modeling Core Metabolism in Cancer Cells: Surveying the Topology Underlying the Warburg Effect. *PLoS One*, 5(8):e12383, August 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0012383. URL http://dx.plos.org/10.1371/journal.pone.0012383.

[60] Tomer Shlomi, Tomer Benyamini, Eyal Gottlieb, Roded Sharan, and Eytan Ruppin. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS Comput. Biol.*, 7(3):e1002018, March 2011. ISSN 1553-7358.

doi: 10.1371/journal.pcbi.1002018. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3053319&tool=pmcentrez&rendertype=abstract.

[61] Alexei Vazquez and Zoltán N Oltvai. Molecular crowding defines a common origin for the Warburg effect in proliferating cells and the lactate threshold in muscle physiology. *PLoS One*, 6(4):e19538, January 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0019538. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3084886&tool=pmcentrez&rendertype=abstract.

[62] Christian Frezza, Liang Zheng, Ori Folger, Kartik N Rajagopalan, Elaine D MacKenzie, Livnat Jerby, Massimo Micaroni, Barbara Chaneton, Julie Adam, Ann Hedley, Gabriela Kalna, Ian P M Tomlinson, Patrick J Pollard, Dave G Watson, Ralph J Deberardinis, Tomer Shlomi, Eytan Ruppin, and Eyal Gottlieb. Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature*, 477(7363):225–8, September 2011. ISSN 1476-4687. doi: 10.1038/nature10363. URL http://www.ncbi.nlm.nih.gov/pubmed/21849978.

[63] Ines Thiele, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, Stefan G Thorleifsson, Rasmus Agren, Christian Bölling, Sergio Bordel, Arvind K Chavali, Paul Dobson, Warwick B Dunn, Lukas Endler, David Hala, Michael Hucka, Duncan Hull, Daniel Jameson, Neema Jamshidi, Jon J Jonsson, Nick Juty, Sarah Keating, Intawat Nookaew, Nicolas Le Novère, Naglis Malys, Alexander Mazein, Jason A Papin, Nathan D Price, Anatoly Sorokin, Johannes H G M Van Beek, Dieter Weichart, Igor Goryanin, and Jens Nielsen. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, 31(March 2013):1–9, 2013. doi: 10.1038/nbt.2488.

[64] Erwin P Gianchandani, Arvind K Chavali, and Jason A Papin. The application of flux balance analysis in systems biology. 2009. doi: 10.1002/wsbm.060.

[65] Nathan E Lewis, Harish Nagarajan, and Bernhard Ø Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.*, 10(4):291–305. ISSN 1740-1534. doi: 10.1038/nrmicro2737. URL http://www.ncbi.nlm.nih.gov/pubmed/22367118.

[66] Vineet Sangar, James a Eddy, Evangelos Simeonidis, and Nathan D Price. Mechanistic modeling of aberrant energy metabolism in human disease. *Front. Physiol.*, 3(October):404, January 2012. ISSN 1664-042X. doi: 10.3389/fphys.2012.00404. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3480659&tool=pmcentrez&rendertype=abstract.

[67] Joshua a Lerman, Daniel R Hyduke, Haythem Latif, Vasiliy a Portnoy, Nathan E Lewis, Jeffrey D Orth, Alexandra C Schrimpe-Rutledge, Richard D Smith, Joshua N Adkins, Karsten Zengler, and Bernhard Ø Palsson. In silico method for modelling metabolism and gene

product expression at genome scale. *Nat. Commun.*, 3(may):929, January 2012. ISSN 2041-1723. doi: 10.1038/ncomms1928. URL http://www.ncbi.nlm.nih.gov/pubmed/22760628.

[68] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, July 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.05.044. URL http://www.ncbi.nlm.nih.gov/pubmed/22817898.

[69] Keren Yizhak, Tomer Benyamini, Wolfram Liebermeister, Eytan Ruppin, and Tomer Shlomi. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26(12):i255–60, June 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq183. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2881368&tool=pmcentrez&rendertype=abstract.

[70] Markus J Herrgård, Neil Swainston, Paul Dobson, Warwick B Dunn, K Yalçin Arga, Mikko Arvas, Nils Blüthgen, Simon Borger, Roeland Costenoble, Matthias Heinemann, Michael Hucka, Nicolas Le Novère, Peter Li, Wolfram Liebermeister, Monica L Mo, Ana Paula Oliveira, Dina Petranovic, Stephen Pettifer, Evangelos Simeonidis, Kieran Smallbone, Irena Spasić, Dieter Weichart, Roger Brent, David S Broomhead, Hans V Westerhoff, Betül Kirdar, Merja Penttilä, Edda Klipp, Bernhard Ø Palsson, Uwe Sauer, Stephen G Oliver, Pedro Mendes, Jens Nielsen, and Douglas B Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, 26(10):1155–60, October 2008. ISSN 1546-1696. doi: 10.1038/nbt1492. URL http://www.ncbi.nlm.nih.gov/pubmed/18846089.

[71] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Abhishek Arora, and Markus W Covert. WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic Acids Res.*, 41(Database issue):D787–92, January 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1108. URL http://www.ncbi.nlm.nih.gov/pubmed/23175606.

[72] http://kbase.science.energy.gov/ (Accessed March 3 2013).

[73] Stephan Pabinger, Robert Rader, Rasmus Agren, Jens Nielsen, and Zlatko Trajanoski. MEMOSys: Bioinformatics platform for genome-scale metabolic models. *BMC Syst. Biol.*, 5(1):20, January 2011. ISSN 1752-0509. doi: 10.1186/1752-0509-5-20. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3045322&tool=pmcentrez&rendertype=abstract.

[74] Tomáš Helikar, Bryan Kowal, Sean McClenathan, Mitchell Bruckner, Thaine Rowley, Alex Madrahimov, Ben Wicks, Manish Shrestha, Kahani Limbu, and Jim a Rogers. The Cell Collective: toward an open and collaborative approach to systems biology. *BMC Syst. Biol.*, 6:96, January 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-96. URL http://www.pubmedcentral.nih.gov/articlerender.

fcgi?artid=3443426&tool=pmcentrez&rendertype=abstract.

[75] Lin Xu. *Dynamics of epistasis from duplicate genes to genome-wide networks*. PhD thesis, Cornell University, Ann Arbor, 2012. URL http://search.proquest.com/docview/1013994495?accountid=10267.

[76] Patrick C Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9 (11):855–867, November 2008. ISSN 1471-0064. doi: 10.1038/nrg2452. URL http://dx.doi.org/10.1038/nrg2452http://www.ncbi.nlm.nih.gov/pubmed/18852697.

[77] Charles Boone, Howard Bussey, and Brenda J Andrews. Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.*, 8(6): 437–49, June 2007. ISSN 1471-0056. doi: 10.1038/nrg2085. URL http://www.ncbi.nlm.nih.gov/pubmed/17510664.

[78] Motoo Kimura and Takeo Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54(6):1337–1351, June 1966. ISSN 00754617. doi: 10.1002/jlac.19666940111. URL http://doi.wiley.com/10.1002/jlac.19666940111.

[79] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice L Y Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P St Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M Wallace, Joseph A Whitney, Matthew T Weirauch, Guoqing Zhong, Hongwei Zhu, Walid A Houry, Michael Brudno, Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P Roth, Guri Giaever, Corey Nislow, Olga G Troyanskaya, Howard Bussey, Gary D Bader, Anne-Claude Gingras, Quaid D Morris, Philip M Kim, Chris A Kaiser, Chad L Myers, Brenda J Andrews, and Charles Boone. The genetic landscape of a cell. *Science*, 327 (5964):425–31, 2010. ISSN 1095-9203. doi: 10.1126/science.1180823. URL http://www.ncbi.nlm.nih.gov/pubmed/20093466.

[80] Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F Berriz, Renee L Brost, Michael Chang, YiQun Chen, Xin Cheng, Gordon Chua, Helena Friesen, Debra S Goldberg, Jennifer Haynes, Christine Humphries, Grace He, Shamiza Hussein, Lizhu Ke, Nevan Krogan, Zhijian Li, Joshua N Levinson, Hong Lu, Patrice Ménard, Christella Munyana, Ainslie B Parsons, Owen Ryan, Raffi Tonikian, Tania Roberts, Anne-Marie Sdicu, Jesse Shapiro, Bilal Sheikh, Bernhard Suter, Sharyl L Wong, Lan V Zhang, Hongwei Zhu, Christopher G Burd, Sean Munro, Chris Sander, Jasper Rine, Jack Greenblatt, Matthias Peter, Anthony Bretscher, Graham Bell, Frederick P Roth, Grant W Brown, Brenda Andrews, Howard Bussey, and Charles Boone. Global Mapping of the Yeast Genetic Interaction Network. *Science (80-.).*, 303(5659):808–813, February 2004. doi: 10.1126/science.1091317. URL http://www.sciencemag.org/content/303/5659/808.abstract.

[81] Xuewen Pan, Daniel S. Yuan, Dong Xiang, Xiaoling Wang, Sharon Sookhai-Mahadeo, Joel S. Bader, Philip Hieter, Forrest Spencer, and Jef D. Boeke. A robust toolkit for functional profiling of the yeast genome. *Mol. Cell*, 16:487–496, 2004. ISSN 10972765. doi: 10.1016/j.molcel.2004.09.035.

[82] Xuewen Pan, Ping Ye, Daniel S. Yuan, Xiaoling Wang, Joel S. Bader, and Jef D. Boeke. A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell*, 124:1069–1081, 2006. ISSN 00928674. doi: 10.1016/j.cell.2005.12.036.

[83] V Measday and P Hieter. Synthetic dosage lethality. In *Guid. TO YEAST Genet. Mol. CELL Biol. PT B*, volume 350 of *METHODS IN ENZYMOLOGY*, pages 316–326. ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495 USA, 2002.

[84] Vivien Measday, Kristin Baetz, Julie Guzzo, Karen Yuen, Teresa Kwok, Bilal Sheikh, Huiming Ding, Ryo Ueta, Trinh Hoac, Benjamin Cheng, Isabelle Pot, Amy Tong, Yuko Yamaguchi-Iwai, Charles Boone, Phil Hieter, and Brenda Andrews. Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc. Natl. Acad. Sci. U. S. A.*, 102:13956–13961, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0503504102.

[85] Richelle Sopko, Dongqing Huang, Nicolle Preston, Gordon Chua, Balázs Papp, Kimberly Kafadar, Mike Snyder, Stephen G. Oliver, Martha Cyert, Timothy R. Hughes, Charles Boone, and Brenda Andrews. Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell*, 21:319–330, 2006. ISSN 10972765. doi: 10.1016/j.molcel.2005.12.011.

[86] Sean R Collins, Kyle M Miller, Nancy L Maas, Assen Roguev, Jeffrey Fillingham, Clement S Chu, Maya Schuldiner, Marinella Gebbia, Judith Recht, Michael Shales, Huiming Ding, Hong Xu, Junhong Han, Kristin Ingvarsdottir, Benjamin Cheng, Brenda Andrews, Charles Boone, Shelley L Berger, Phil Hieter, Zhiguo Zhang, Grant W Brown, C James Ingles, Andrew Emili, C David Allis, David P Toczyski, Jonathan S Weissman, Jack F Greenblatt, and Nevan J Krogan. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, 446:806–810, 2007. ISSN 0028-0836. doi: 10.1038/nature05649.

[87] Benoît Kornmann, Erin Currie, Sean R Collins, Maya Schuldiner, Jodi Nunnari, Jonathan S Weissman, and Peter Walter. An ER-mitochondria tethering complex revealed by a synthetic biology screen. *Science*, 325:477–481, 2009. ISSN 0036-8075. doi: 10.1126/science.1175088.

[88] Dorothea Fiedler, Hannes Braberg, Monika Mehta, Gal Chechik, Gerard Cagney, Paromita Mukherjee, Andrea C Silva, Michael Shales, Sean R Collins, Sake van Wageningen, Patrick Kemmeren, Frank C P Holstege, Jonathan S Weissman, Michael-Christopher Keogh, Daphne Koller, Kevan M Shokat, and Nevan J Krogan. Functional organization of the S. cerevisiae phosphorylation network. *Cell*, 136(5):952–63, 2009. ISSN 1097-4172. doi: 10.1016/j.cell.2008.12.039. URL `http://www.ncbi.nlm.nih.gov/pubmed/19269370`.

[89] Sebastian Bonhoeffer, Colombe Chappey, Neil T Parkin, Jeanette M Whitcomb, and Christos J Petropoulos. Evidence for positive epistasis in HIV-1. *Science*, 306:1547–1550, 2004. ISSN 0036-8075. doi: 10.1126/science.1101786.

[90] Assen Roguev, Sourav Bandyopadhyay, Martin Zofall, Ke Zhang, Tamas Fischer, Sean R Collins, Hongjing Qu, Michael Shales, Han-Oh Park, Jacqueline Hayles, Kwang-Lae Hoe, Dong-Uk Kim, Trey Ideker, Shiv I Grewal, Jonathan S Weissman, and Nevan J Krogan. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, 322:405–410, 2008. ISSN 0036-8075. doi: 10.1126/science.1162609.

[91] Monica L Mo, Bernhard Ø Palsson, and Markus J Herrgård. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.*, 3(1):37, 2009. ISSN 1752-0509. doi: 10.1186/1752-0509-3-37. URL http://www.biomedcentral.com/1752-0509/3/37.

[92] Scott A Becker, Adam M Feist, Monica L Mo, Gregory Hannum, Bernhard Ø Palsson, and Markus J Herrgård. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat. Protoc.*, 2(3):727–38, 2007. ISSN 1750-2799. doi: 10.1038/nprot.2007.99. URL http://www.ncbi.nlm.nih.gov/pubmed/17406635.

[93] Balázs Papp, Csaba Pál, and Laurence D Hurst. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992):661–4, June 2004. ISSN 1476-4687. doi: 10.1038/nature02636. URL http://www.ncbi.nlm.nih.gov/pubmed/15190353.

[94] Rafael U Ibarra, Jeremy S Edwards, and Bernhard Ø Palsson. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912):186–189, November 2002. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature01149.

[95] Richard Harrison, Balázs Papp, Csaba Pál, Stephen G Oliver, and Daniela Delneri. Plasticity of genetic interactions in metabolic networks of yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 104(7):2307–12, February 2007. ISSN 0027-8424. doi: 10.1073/pnas.0607153104. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1892960&tool=pmcentrez&rendertype=abstract.

[96] David Deutscher, Isaac Meilijson, Martin Kupiec, and Eytan Ruppin. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat. Genet.*, 38(9):993–8, September 2006. ISSN 1061-4036. doi: 10.1038/ng1856. URL http://www.ncbi.nlm.nih.gov/pubmed/16941010.

[97] Daniel Segrè, Alexander Deluna, George M Church, and Roy Kishony. Modular epistasis in yeast metabolism. *Nat. Genet.*, 37(1):77–83, January 2005. ISSN 1061-4036. doi: 10.1038/ng1489. URL http://www.ncbi.nlm.nih.gov/pubmed/15592468.

[98] J S Edwards, R U Ibarra, and B Ø Palsson. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, 19:125–130, 2001. ISSN 1087-0156. doi: 10.1038/84379.

[99] Manal AbuOun, Patrick F Suthers, Gareth I Jones, Ben R Carter, Mark P Saunders, Costas D Maranas, Martin J Woodward, and Muna F Anjum. Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal Escherichia coli strain. *J. Biol. Chem.*, 284:29480–29488, 2009. ISSN 0021-9258. doi: 10.1074/jbc.M109.005868.

[100] Maxime Durot, Pierre-Yves Bourguignon, and Vincent Schachter. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol. Rev.*, 33:164–190, 2009. ISSN 0168-6445. doi: 10.1111/j.1574-6976.2008.00146.x.

[101] Adam M Feist and Bernhard Ø Palsson. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.*, 26:659–667, 2008. ISSN 1087-0156. doi: 10.1038/nbt1401.

[102] Stephen S Fong, Anthony P Burgard, Christopher D Herring, Eric M Knight, Frederick R Blattner, Costas D Maranas, and Bernhard Ø Palsson. In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnol. Bioeng.*, 91:643–648, 2005. ISSN 0006-3592. doi: 10.1002/bit.20542.

[103] Stephen S Fong, Andrew R Joyce, and Bernhard Ø Palsson. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Res.*, 15(10):1365–72, October 2005. ISSN 1088-9051. doi: 10.1101/gr.3832305. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1240078&tool=pmcentrez&rendertype=abstract.

[104] Christina L Burch and Lin Chao. Epistasis and its relationship to canalization in the RNA virus phi 6. *Genetics*, 167:559–567, 2004. ISSN 0016-6731. doi: 10.1534/genetics.103.021196.

[105] Lingchong You and John Yin. Dependence of epistasis on environment and mutation severity as revealed by in silico mutagenesis of phage t7. *Genetics*, 160:1273–1281, 2002. ISSN 0016-6731.

[106] Rafael Sanjuán. Quantifying antagonistic epistasis in a multifunctional RNA secondary structure of the Rous sarcoma virus. *J. Gen. Virol.*,

87:1595–1602, 2006. ISSN 0022-1317. doi: 10.1099/vir.0.81585-0.

[107] Ricardo B R Azevedo, Rolf Lohaus, Suraj Srinivasan, Kristen K Dang, and Christina L Burch. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature*, 440(7080):87–90, March 2006. ISSN 1476-4687. doi: 10.1038/nature04488. URL http://www.ncbi.nlm.nih.gov/pubmed/16511495.

[108] Rolf Lohaus, Christina L Burch, and Ricardo B R Azevedo. Genetic architecture and the evolution of sex. *J. Hered.*, 101 Suppl:S142–57, 2010. ISSN 1465-7333. doi: 10.1093/jhered/esq013. URL http://www.ncbi.nlm.nih.gov/pubmed/20421324.

[109] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard Ø Palsson. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, 3(121):121, January 2007. ISSN 1744-4292. doi: 10.1038/msb4100155. URL http://www.ncbi.nlm.nih.gov/pubmed/17593909.

[110] Ines Thiele, Daniel R Hyduke, Benjamin Steeb, Guy Fankam, Douglas K Allen, Susanna Bazzani, Pep Charusanti, Feng-Chi Chen, Ronan M T Fleming, Chao A Hsiung, Sigrid C J De Keersmaecker, Yu-Chieh Liao, Kathleen Marchal, Monica L Mo, Emre Özdemir, Anu Raghunathan, Jennifer L Reed, Sook-il Shin, Sara Sigurbjörnsdóttir, Jonas Steinmann, Suresh Sudarsan, Neil Swainston, Inge M Thijs, Karsten Zengler, Bernhard Ø Palsson, Joshua N Adkins, and Dirk Bumann. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Syst. Biol.*, 5:8, 2011. ISSN 1752-0509. doi: 10.1186/1752-0509-5-8.

[111] Ines Thiele, Thuy D Vo, Nathan D Price, and Bernhard Ø Palsson. Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *J. Bacteriol.*, 187:5818–5830, 2005. ISSN 0021-9193. doi: 10.1128/JB.187.16.5818-5830.2005.

[112] Adam M Feist, Johannes C M Scholten, Bernhard Ø Palsson, Fred J Brockman, and Trey Ideker. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol. Syst. Biol.*, 2:2006.0004, 2006. ISSN 1744-4292. doi: 10.1038/msb4100046.

[113] Germán Plata, Tzu-Lin Hsiao, Kellen L Olszewski, Manuel Llinás, and Dennis Vitkup. Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network. *Mol. Syst. Biol.*, 6:408, 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.60.

[114] Rafael Sanjuán and Santiago F Elena. Epistasis correlates to genomic complexity. *Proc. Natl. Acad. Sci. U. S. A.*, 103(39):14402–5, September 2006. ISSN 0027-8424. doi: 10.1073/pnas.0604543103. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1599975&tool=pmcentrez&rendertype=abstract.

[115] Rafael Sanjuán and Miguel R. Nebot. A network model for the correlation between epistasis and genomic complexity. *PLoS One*, 3, 2008. ISSN 19326203. doi: 10.1371/journal.pone.0002663.

[116] A S Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336:435–440, 1988. ISSN 0028-0836. doi: 10.1038/336435a0.

[117] Sarah P Otto. Unravelling the evolutionary advantage of sex: a commentary on 'Mutation-selection balance and the evolutionary advantage of sex and recombination' by Brian Charlesworth. *Genet. Res.*, 89(5-6):447–9, December 2007. ISSN 1469-5073. doi: 10.1017/S001667230800966X. URL http://www.ncbi.nlm.nih.gov/pubmed/18976534.

[118] J Arjan G M de Visser and Santiago F Elena. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat. Rev. Genet.*, 8:139–149, 2007. ISSN 1471-0056. doi: 10.1038/nrg1985.

[119] Roger D Kouyos, Olin K Silander, and Sebastian Bonhoeffer. Epistasis between deleterious mutations and the evolution of recombination. *Trends Ecol. Evol.*, 22(6):308–15, June 2007. ISSN 0169-5347. doi: 10.1016/j.tree.2007.02.014. URL http://www.ncbi.nlm.nih.gov/pubmed/17337087.

[120] Balázs Szappanos, Károly Kovács, Béla Szamecz, Frantisek Honti, Michael Costanzo, Anastasia Baryshnikova, Gabriel Gelius-Dietrich, Martin J Lercher, Márk Jelasity, Chad L Myers, Brenda J Andrews, Charles Boone, Stephen G Oliver, Csaba Pál, and Balázs Papp. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.*, 43(May):656–662, May 2011. ISSN 1546-1718. doi: 10.1038/ng.846. URL http://www.ncbi.nlm.nih.gov/pubmed/21623372.

[121] Scott Banta, Murali Vemula, Tadaaki Yokoyama, Arul Jayaraman, François Berthiaume, and Martin L Yarmush. Contribution of gene expression to metabolic fluxes in hypermetabolic livers induced through burn injury and cecal ligation and puncture in rats. *Biotechnol. Bioeng.*, 97:118–137, 2007. ISSN 1559-047X. doi: 10.1097/01253092-200603001-00230.

[122] Daniel L Hartl and Andrew G Clark. *Principles of Population Genetics*. Sinauer Associates, Inc., Sinauer, Sunderland, MA, 4th edition, 2007. ISBN 0878933085.

[123] Alexey S Kondrashov. Selection against harmful mutations in large sexual and asexual populations. *Genet. Res. (Camb).*, 40(03):325–332, 1982.

[124] Daven C Presgraves. Speciation Genetics: Epistasis, Conflict and the Origin of Species. *Curr. Biol.*, 17(4):R125–R127, July 2007. doi: 10.1016/j.cub.2006.12.030. URL http://www.cell.com/current-biology/abstract/S0960-9822(06)02664-9.

[125] Thomas F Hansen and Günter P Wagner. Epistasis and the Mutation Load: A Measurement-Theoretical Approach. *Genet.*, 158(1):477–485, May 2001. URL http://www.genetics.org/content/158/1/477.abstract.

[126] Gabriel Musso, Michael Costanzo, ManQin Huangfu, Andrew M Smith, Jadine Paw, Bryan-Joseph San Luis, Charles Boone, Guri Giaever, Corey Nislow, Andrew Emili, and Zhaolei Zhang. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.*, 18(7):1092–1099, July 2008. doi: 10.1101/gr.076174.108. URL http://genome.cshlp.org/content/18/7/1092.abstract.

[127] Lin Xu, Huifeng Jiang, Hong Chen, and Zhenglong Gu. Genetic Architecture of Growth Traits Revealed by Global Epistatic Interactions. *Genome Biol. Evol.*, 3:909–914, January 2011. doi: 10.1093/gbe/evr065. URL http://gbe.oxfordjournals.org/content/3/909.abstract.

[128] Andrés Pérez-Figueroa, Armando Caballero, Aurora García-Dorado, and Carlos López-Fanjul. The Action of Purifying Selection, Mutation and Drift on Fitness Epistatic Systems. *Genet.*, 183(1):299–313, September 2009. doi: 10.1534/genetics.109.104893. URL http://www.genetics.org/content/183/1/299.abstract.

[129] Sandra Trindade, Ana Sousa, Karina Bivar Xavier, Francisco Dionisio, Miguel Godinho Ferreira, and Isabel Gordo. Positive Epistasis Drives the Acquisition of Multidrug Resistance. *PLoS Genet*, 5(7):e1000578, July 2009. URL http://dx.doi.org/10.1371/journal.pgen.1000578.

[130] Susanna K Remold and Richard E Lenski. Contribution of individual random mutations to genotype-by-environment interactions in Escherichia coli. *Proc. Natl. Acad. Sci.*, 98(20):11388–11393, September 2001. doi: 10.1073/pnas.201140198. URL http://www.pnas.org/content/98/20/11388.abstract.

[131] Roy Kishony and Stanislas Leibler. Environmental stresses can alleviate the average deleterious effect of mutations. *J. Biol.*, 2(2):14, 2003. ISSN 1475-4924. URL http://jbiol.com/content/2/2/14.

[132] Tim F Cooper, Richard E Lenski, and Santiago F Elena. Parasites and mutational load: an experimental test of a pluralistic theory for the evolution of sex. *Proc. Biol. Sci.*, 272:311–317, 2005. ISSN 0962-8452. doi: 10.1098/rspb.2004.2975.

[133] Ryszard Korona. Genetic Load of the Yeast Saccharomyces cerevisiae under Diverse Environmental Conditions. *Evolution (N. Y).*, 53(6): 1966–1971, December 1999. ISSN 00143820. doi: 10.2307/2640455. URL http://www.jstor.org/stable/2640455.

[134] Krzysztof Szafraniec, Rhona H Borts, and Ryszard Korona. Environmental stress and mutational load in diploid strains of the yeast Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.*, 98(3):1107–1112, January 2001. doi: 10.1073/pnas.98.3.1107. URL http://www.pnas.org/content/98/3/1107.abstract.

[135] Lukasz Jasnos, Katarzyna Tomala, Dorota Paczesniak, and Ryszard Korona. Interactions Between Stressful Environment and Gene Deletions Alleviate the Expected Average Loss of Fitness in Yeast. *Genet.*, 178(4):2105–2111, April 2008. doi: 10.1534/genetics.107.084533. URL http://www.genetics.org/content/178/4/2105.abstract.

[136] Larissa L Vassilieva, Aaron M Hook, and Michael Lynch. THE FITNESS EFFECTS OF SPONTANEOUS MUTATIONS IN CAENORHABDITIS ELEGANS. *Evolution (N. Y).*, 54(4):1234–1246, August 2000. ISSN 1558-5646. doi: 10.1111/j.0014-3820.2000.tb00557.x. URL http://dx.doi.org/10.1111/j.0014-3820.2000.tb00557.x.

[137] Charles F Baer, Naomi Phillips, Dejerianne Ostrow, Arián Avalos, Dustin Blanton, Ashley Boggs, Thomas Keller, Laura Levy, and Edward Mezerhane. Cumulative Effects of Spontaneous Mutations for Fitness in Caenorhabditis: Role of Genotype, Environment and Stress. *Genet.*, 174(3):1387–1395, November 2006. doi: 10.1534/genetics.106.061200. URL http://www.genetics.org/content/174/3/1387.abstract.

[138] Hsiao-Pei Yang, Ana Y Tanikawa, Wayne A Van Voorhies, Joana C Silva, and Alexey S Kondrashov. Whole-Genome Effects of Ethyl Methanesulfonate-Induced Mutation on Nine Quantitative Traits in Outbred Drosophila melanogaster. *Genet.*, 157(3):1257–1265, March 2001. URL http://www.genetics.org/content/157/3/1257.abstract.

[139] James D Fry and Stefanie L Heinsohn. Environment Dependence of Mutational Parameters for Viability in Drosophila melanogaster. *Genet.*, 161(3):1155–1167, July 2002. URL http://www.genetics.org/content/161/3/1155.abstract.

[140] Alethea D Wang, Nathaniel P Sharp, Christine C Spencer, Katherine Tedman-Aucoin, and Aneil F Agrawal. Selection, Epistasis, and ParentofOrigin Effects on Deleterious Mutations across Environments in Drosophila melanogaster. *Am. Nat.*, 174(6):863–874, December

2009. ISSN 00030147. doi: 10.1086/645088. URL http://www.jstor.org/stable/10.1086/645088.

[141] Jadene A Young, Christopher P Yourth, and Aneil F Agrawal. The effect of pathogens on selection against deleterious mutations in Drosophila melanogaster. *J. Evol. Biol.*, 22(10):2125–2129, October 2009. ISSN 1420-9101. doi: 10.1111/j.1420-9101.2009.01830.x. URL http://dx.doi.org/10.1111/j.1420-9101.2009.01830.x.

[142] Sourav Bandyopadhyay. Rewiring of Genetic Networks in Response to DNA Damage. *Science (80-.).*, 1385(2010), 2011. doi: 10.1126/science.1195618.

[143] Agata Jakubowska and Ryszard Korona. Epistasis for Growth Rate and Total Metabolic Flux in Yeast. *PLoS One*, 7(3):e33132, March 2012. URL http://dx.doi.org/10.1371/journal.pone.0033132.

[144] Dennis P Wall, Aaron E Hirsh, Hunter B Fraser, Jochen Kumm, Guri Giaever, Michael B Eisen, and Marcus W Feldman. Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. United States Am.*, 102(15):5483–5488, April 2005. doi: 10.1073/pnas.0501761102. URL http://www.pnas.org/content/102/15/5483.abstract.

[145] Ben Lehner. Molecular mechanisms of epistasis within and between genes. *Trends Genet.*, 27(8):323–331, July 2011. doi: 10.1016/j.tig.2011.05.007. URL http://www.cell.com/trends/genetics/abstract/S0168-9525(11)00077-1.

[146] Aneil F Agrawal and Michael C Whitlock. Environmental duress and epistasis: how does stress affect the strength of selection on new mutations? *Trends Ecol. Evol.*, 25(8):450–8, August 2010. ISSN 0169-5347. doi: 10.1016/j.tree.2010.05.003. URL http://www.ncbi.nlm.nih.gov/pubmed/20538366.

[147] M Olivia Casanueva, Alejandro Burga, and Ben Lehner. Fitness Trade-Offs and Environmentally Induced Mutation Buffering in Isogenic C. elegans. *Sci.*, 335(6064):82–85, January 2012. doi: 10.1126/science.1213491. URL http://www.sciencemag.org/content/335/6064/82.abstract.

[148] Anthony P Burgard, Priti Pharkya, and Costas D Maranas. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.*, 84(6):647–657, December 2003. ISSN 1097-0290. doi: 10.1002/bit.10803. URL http://dx.doi.org/10.1002/bit.10803.

[149] Kiran Patil, Isabel Rocha, Jochen Forster, and Jens Nielsen. Evolutionary programming as a platform for in silico metabolic engineering.

BMC Bioinformatics, 6(1):308, 2005. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/6/308.

[150] Sean P. Cornelius, William L. Kath, and Adilson E. Motter. Controlling Complex Networks with Compensatory Perturbations. May 2011. URL http://arxiv.org/abs/1105.3726.

[151] Ana Rita Brochado, Sergej Andrejev, Costas D Maranas, and Kiran R Patil. Impact of Stoichiometry Representation on Simulation of Genotype-Phenotype Relationships in Metabolic Networks. *PLoS Comput Biol*, 8(11):e1002758, November 2012. URL http://dx.doi.org/10.1371/journal.pcbi.1002758.

[152] Jan Schellenberger, Richard Que, Ronan M T Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, Joseph Kang, Daniel R Hyduke, and Bernhard O Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.*, 6(9):1290–1307, September 2011. ISSN 1754-2189. URL http://dx.doi.org/10.1038/nprot.2011.308http://www.nature.com/nprot/journal/v6/n9/abs/nprot.2011.308.html#supplementary-information.

[153] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2): 101–113, February 2004. ISSN 1471-0056. URL http://dx.doi.org/10.1038/nrg1272.

[154] Nathan E Lewis, Harish Nagarajan, and Bernhard Ø Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.*, 10(4):291–305, April 2012. ISSN 1740-1534. doi: 10.1038/nrmicro2737. URL http://www.ncbi.nlm.nih.gov/pubmed/22367118.

[155] Sergej Pirkmajer and Alexander V Chibalin. Serum starvation: caveat emptor. *Am. J. Physiol. Cell Physiol.*, 301(2):C272–C279, July 2011. URL http://ajpcell.physiology.org/highwire/citation/1105/mendeley.

[156] Anna S Blazier and Jason a Papin. Integration of expression data in genome-scale metabolic network reconstructions. *Front. Physiol.*, 3(August):299, January 2012. ISSN 1664-042X. doi: 10.3389/fphys.2012.00299. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3429070&tool=pmcentrez&rendertype=abstract.

[157] Daniel Segrè, Dennis Vitkup, and George M Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.*, 99(23):15112–7, November 2002. ISSN 0027-8424. doi: 10.1073/pnas.232349399. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=137552&tool=pmcentrez&rendertype=abstract.

[158] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 2004. ISBN 978-0-521-83378-3. URL http://www.stanford.edu/~boyd/cvxbook/.

[159] Christoph Jüschke, Ilse Dohnal, Peter Pichler, Heike Harzer, Remco Swart, Gustav Ammerer, Karl Mechtler, and Juergen A Knoblich. Transcriptome and proteome quantification of a tumor model provides novel insights into post-transcriptional gene regulation. *Genome Biol.*, 14(11):1–33, 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-11-r133. URL http://genomebiology.com/2013/14/11/R133.

[160] Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach, 3rd edition. *Prentice Hall*, 2009. ISSN 0269-8889. doi: 10.1017/S0269888900007724. URL http://portal.acm.org/citation.cfm?id=1671238&coll=DL&dl=GUIDE&CFID=19086450 1&CFTOKEN=29051579$\delimiter"026E30F$npapers2://publication/uuid/4B787E16-89F6-4FF7-A5E5-E59F3CFEFE88.

[161] Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua a Lerman, Hojung Nam, Adam M Feist, and Bernhard Ø Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism2011. *Mol. Syst. Biol.*, 7(535):1–9, October 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.65. URL http://www.nature.com/doifinder/10.1038/msb.2011.65.

[162] Hnin W Aung, Susan A Henry, and Larry P Walker. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Ind. Biotechnol.*, 9(4):215–228, August 2013. ISSN 1550-9087. doi: 10.1089/ind.2013.0013. URL http://online.liebertpub.com/doi/abs/10.1089/ind.2013.0013.

[163] Paola Picotti, Mathieu Clement-Ziza, Henry Lam, David S Campbell, Alexander Schmidt, Eric W Deutsch, Hannes Rost, Zhi Sun, Oliver Rinner, Lukas Reiter, Qin Shen, Jacob J Michaelson, Andreas Frei, Simon Alberti, Ulrike Kusebauch, Bernd Wollscheid, Robert L Moritz, Andreas Beyer, and Ruedi Aebersold. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, 494(7436):266–270, February 2013. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature11835http://www.nature.com/nature/journal/v494/n7436/abs/nature11835.html#supplementary-information.

[164] Amin Moghaddas Gholami, Hannes Hahne, Zhixiang Wu, Florian Johann Auer, Chen Meng, Mathias Wilhelm, and Bernhard Kuster. Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Rep.*, 4(3):609–620, August 2013. URL http://linkinghub.elsevier.com/retrieve/pii/S221112471300380X.

[165] Benjamin D Heavner, Kieran Smallbone, Brandon Barker, Pedro Mendes, and Larry P Walker. Yeast 5 - an expanded reconstruction of the Saccharomyces cerevisiae metabolic network. *BMC Syst. Biol.*, 6(1):55, January 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-55. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3413506&tool=pmcentrez&rendertype=abstract.

[166] Bryson D Bennett, Elizabeth H Kimball, Melissa Gao, Robin Osterhout, Stephen J Van Dien, and Joshua D Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat. Chem. Biol.*, 5(8):593–599, August 2009. ISSN 1552-4450. URL http://dx.doi.org/10.1038/nchembio.186http://www.nature.com/nchembio/journal/v5/n8/suppinfo/nchembio.186_S1.html.

[167] Victor Chubukov, Markus Uhr, Ludovic Le Chat, Roelco J Kleijn, Matthieu Jules, Hannes Link, Stephane Aymerich, Jorg Stelling, and Uwe Sauer. Transcriptional regulation is insufficient to explain substrate-induced flux changes in Bacillus subtilis. *Mol. Syst. Biol.*, 9, November 2013. URL http://dx.doi.org/10.1038/msb.2013.6610.1038/msb.2013.66.

[168] Jennifer L Reed, Trina R Patel, Keri H Chen, Andrew R Joyce, Margaret K Applebee, Christopher D Herring, Olivia T Bui, Eric M Knight, Stephen S Fong, and Bernhard Ø Palsson. Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci.*, 103(46):17480–17484, 2006. doi: 10.1073/pnas.0603364103. URL http://www.pnas.org/content/103/46/17480.abstract.

[169] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.*, 14(10):719–732, October 2013. ISSN 1471-0056. URL http://dx.doi.org/10.1038/nrg355210.1038/nrg3552http://www.nature.com/nrg/journal/v14/n10/abs/nrg3552.html#supplementary-information.

[170] Germán Plata, Tobias Fuhrer, Tzu-Lin Hsiao, Uwe Sauer, and Dennis Vitkup. Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nat. Chem. Biol.*, 8(10):848–854, October 2012. ISSN 1552-4450. URL http://dx.doi.org/10.1038/nchembio.1063http://www.nature.com/nchembio/journal/v8/n10/abs/nchembio.1063.html#supplementary-information.

[171] Motoo Kimura. The neutral theory of molecular evolution and the world view of the neutralists. *Genome*, 31(1):24–31, January 1989. ISSN 0831-2796. doi: 10.1139/g89-009. URL http://dx.doi.org/10.1139/g89-009.

[172] H Allen Orr. The Distribution of Fitness Effects Among Beneficial Mutations. *Genetics*, 163(4):1519–1526, April 2003. URL http://www.genetics.org/content/163/4/1519.abstract.

[173] Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F Delaney, Daniel Segrè, and Christopher J Marx. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, 332(6034):1190–2, June 2011. ISSN 1095-9203. doi: 10.1126/science.1203799. URL http://www.ncbi.nlm.nih.gov/pubmed/21636771.

[174] Daniel M Weinreich, Nigel F Delaney, Mark A Depristo, and Daniel L Hartl. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science (80-.).*, 312(7):111–114, 2006.

[175] Zhenyu Xu, Wu Wei, Julien Gagneur, Fabiana Perocchi, Sandra Clauder-Munster, Jurgi Camblong, Elisa Guffanti, Francoise Stutz, Wolfgang Huber, and Lars M Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, February 2009. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature07728http://www.nature.com/nature/journal/v457/n7232/suppinfo/nature07728_S1.html.

[176] Joo Sang Lee, Takashi Nishikawa, and Adilson E Motter. Why optimal states recruit fewer reactions in metabolic networks. *Discret. Contin. Dyn. Syst. - Ser. A*, 32(8):2937 – 2950, 2012. doi: 10.3934/dcds.2012.32.2937.

[177] Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, January 2007. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature05451.

[178] Aisha I Khan, Duy M Dinh, Dominique Schneider, Richard E Lenski, and Tim F Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–6, June 2011. ISSN 1095-9203. doi: 10.1126/science.1203801. URL http://www.ncbi.nlm.nih.gov/pubmed/21636772.

[179] Mark Lunzer, Stephen P Miller, Roderick Felsheim, and Antony M Dean. The Biochemical Architecture of an Ancient Adaptive Landscape. *Science (80-.).*, 310(5747):499–501, October 2005. doi: 10.1126/science.1115649. URL http://www.sciencemag.org/content/310/5747/499.abstract.

[180] H Allen Orr. ADAPTATION AND THE COST OF COMPLEXITY. *Evolution (N. Y).*, 54(1):13–20, February 2000. ISSN 1558-5646. doi: 10.1111/j.0014-3820.2000.tb00002.x. URL http://dx.doi.org/10.1111/j.0014-3820.2000.tb00002.x.

[181] Guillaume Martin, Santiago F Elena, and Thomas Lenormand. Distributions of epistasis in microbes fit predictions from a fitness landscape model. *Nat. Genet.*, 39(4):555–560, April 2007. ISSN 1061-4036. doi: 10.1038/ng1998. URL http://www.ncbi.nlm.nih.gov/pubmed/17369829.

[182] H Allen Orr. THE GENETIC THEORY OF ADAPTATION : A BRIEF HISTORY. *Nat. Rev. Genet.*, 6(February):119–127, 2005. doi: 10.1038/nrg1523.

[183] Guillaume Martin. Fisher's Geometrical Model Emerges as a Property of Complex Integrated Phenotypic Networks. *Genet.*, February 2014. doi: 10.1534/genetics.113.160325. URL http://www.genetics.org/content/early/2014/02/28/genetics.113.160325.abstract.

[184] Jonathan Monk. Available predictive genome-scale metabolic network reconstructions. URL http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms.

[185] Stephen S Fong and Bernhard Ø Palsson. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat. Genet.*, 36(10):1056–8, October 2004. ISSN 1061-4036. doi: 10.1038/ng1432. URL http://www.ncbi.nlm.nih.gov/pubmed/15448692.

[186] Esmaeil Mehrara, Eva Forssell-Aronsson, Håkan Ahlman, and Peter Bernhardt. Quantitative analysis of tumor growth rate and changes in tumor marker level: Specific growth rate versus doubling time. *Acta Oncol. (Madr).*, 48(4):591–597, January 2009. ISSN 0284-186X. doi: 10.1080/02841860802616736. URL http://dx.doi.org/10.1080/02841860802616736.

[187] Edward J O'Brien, Joshua A Lerman, Roger L Chang, Daniel R Hyduke, and Bernhard Ø Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.*, 9, October 2013. URL http://dx.doi.org/10.1038/msb.2013.5210.1038/msb.2013.52.

[188] Brandon Barker, Narayanan Sadagopan, Yiping Wang, Kieran Smallbone, Myers R Christopher, Hongwei Xi, Jason W Locasale, and Zhenglong Gu. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. URL http://arxiv.org/abs/1404.4755.

[189] H Allen Orr. The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. *J. Theor. Biol.*, 238(2):279–85, January 2006. ISSN 0022-5193. doi: 10.1016/j.jtbi.2005.05.001. URL http://www.ncbi.nlm.nih.gov/pubmed/15990119.

[190] Nagarjuna Nagaraj, Jacek R Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Paabo, and Matthias Mann. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.*, 7, November 2011. URL http://dx.doi.org/10.1038/msb.2011.81http://www.nature.com/msb/journal/v7/n1/suppinfo/msb201181_S1.html.

[191] George Oster and Hongyun Wang. Rotary protein motors. *Trends Cell Biol.*, 13(3):114–121, March 2003. URL http://linkinghub.elsevier.com/retrieve/pii/S0962892403000047.

[192] Ida Schomburg, Antje Chang, Sandra Placzek, Carola Söhngen, Michael Rother, Maren Lang, Cornelia Munaretto, Susanne Ulas, Michael Stelzer, Andreas Grote, Maurice Scheer, and Dietmar Schomburg. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.*, 41(D1):D764–D772, January 2013. doi: 10.1093/nar/gks1049. URL http://nar.oxfordjournals.org/content/41/D1/D764.abstract.

[193] David L Nelson and Michael M Cox. *Lehninger Principles of Biochemistry*. W.H. Freeman, New York, NY, USA, 5 edition, 2008. ISBN 071677108X. URL http://www.amazon.com/Lehninger-Principles-Biochemistry-Fourth-Nelson/dp/0716743396.

[194] Gurobi Optimization Inc. Gurobi Optimizer Reference Manual, 2013. URL http://www.gurobi.com.